Linguistic Typology at the Crossroads

ISSN 2785-0943
Volume 5 – Issue 2 – 2025
Issue DOI: 10.60923/issn.2785-0943/v5-n2-2025

Publisher
Department of Classical Philology and Italian Studies (University of Bologna)
Department of Modern Languages, Literatures and Cultures (University of Bologna)
The journal is hosted and maintained by **AlmaDL**

Linguistic Typology at the Crossroads

Responsible Editor

Caterina Mauri, University of Bologna Department of Modern Languages, Literatures and Cultures, via Cartoleria 5, 40124 Bologna. Email: caterina.mauri@unibo.it

Linguistic Typology at the Crossroads

CONTENTS

Linguistic Typology at the Crossroads

Linguistic Typology at the Crossroads 5-2 (2025): I-VII ISSN 2785-0943

At the crossroads of typology and language(s) in use

SILVIA BALLARÈ[1], SIMONE MATTIOLA[2], CATERINA MAURI[1]
[1]ALMA MATER STUDIORUM - UNIVERSITY OF BOLOGNA, [2]UNIVERSITY OF PAVIA

Published: 25/02/2026

A long-standing issue in linguistic typology concerns the relationship between cross-linguistic generalization and the empirical foundations on which such generalizations are built. Typological research necessarily relies on abstraction: comparative concepts are designed to enable systematic comparison across languages while remaining independent from language-particular descriptive categories (Croft 2001; Haspelmath 2010). At the same time, an increasing number of typological studies draw on *naturally occurring data*: typology has increasingly aligned itself with usage-based and corpus-informed perspectives, motivated by the recognition that grammatical structure is shaped by frequency, discourse function, interactional context, and social embedding (Biber 1995; Bybee 2010; Levinson 2016). The resulting methodological and epistemological question is not whether abstraction is required (typology cannot proceed without it) but how abstractions should be calibrated when linguistic structures are examined in situated use.

Variationist sociolinguistics is one of the branches of linguistics that has given more attention to language in use (also from a methodological perspective, see Tagliamonte 2006 inter al.). However, these strands of research have so far only sporadically been integrated in a systematic manner (e.g. Kortmann 2003, Trudgill 2011, Ballarè & Inglese 2023, Sinnemäki 2020), despite the well-established observation that “the patterns of variation and change found in […] a particular language are in many cases simply instances of patterns of variation and change found across languages” (Croft 2022: 27).

https://doi.org/10.60923/issn.2785-0943/24199

The contributions assembled in this issue of *Linguistic Typology at the Crossroads* approach this question from complementary empirical and theoretical perspectives. On the one hand, several papers pursue typological analysis that is explicitly anchored in naturally occurring data, examining how cross-linguistic generalizations are affected when categories and comparisons are grounded in attested usage. On the other hand, the issue includes studies of intralinguistic variation—sometimes drawing on variationist sociolinguistic frameworks—that take variation itself as an empirical window onto grammatical organization; patterns of variation are used to formulate and test generalizations that bear on typological comparison and explanatory modeling. Together, these perspectives invite reflection on how typological generalizations are constructed, evaluated, and interpreted when both cross-linguistic and intralinguistic evidence are incorporated.

Typological generalization and usage-sensitive evidence. Several papers in this issue revisit classic typological domains (such as clause combining, mood and modality, grammatical alternations, and (morpho)syntactic complexity) through systematic analyses of naturally occurring data. A recurring outcome is that patterns often treated as categorical in broad typological overviews display gradedness and probabilistic conditioning once their distribution across comparable usage contexts is examined. This difference is not merely a matter of data quantity, but of evidential type: naturally occurring data foreground optionality, competition among constructions, and skewed frequency profiles that remain largely invisible in elicited or decontextualized data.

These findings sharpen a familiar tension in typology. Explanatory adequacy requires abstraction, yet empirical adequacy requires sensitivity to how forms and functions cluster in use. If typological categories are defined independently of usage conditions, cross-linguistic comparison risks aligning formally similar labels while overlooking functionally distinct distributions; conversely, if comparison is reduced to local distributional profiles, typology risks losing the generality it aims to achieve. The papers in this issue point toward a productive middle position: comparative concepts remain indispensable, but they benefit from being distributionally grounded, that is, tied to attestation conditions, contextual predictors, and frequency-sensitive diagnostics.

From this perspective, frequency effects, register differentiation, and asymmetries in optionality are not noise to be controlled away, but part of what typology must

explain. Usage-based regularities can illuminate why some constructions become cross-linguistically robust, why others remain marginal, and how functional pressures and processing constraints shape grammatical systems over time (cf. Hawkins 2004; Bybee 2010).

Typology, evidence, and the status of explanation. At a more general level, the discussion raised by the contributions in this issue bears directly on the status of explanation in linguistic typology. Typological explanation has traditionally relied on identifying recurrent cross-linguistic patterns and relating them to functional, cognitive, diachronic, or areal factors. While the distinction between descriptive categories and comparative concepts remains fundamental, the nature of the empirical evidence supporting typological explanations has often remained implicit.

The increasing availability of naturally occurring data brings this issue to the foreground, particularly in approaches that model linguistic structure as emerging from usage and distributional regularities (see Du Bois 1985; Levshina 2022). When typological claims are supported by frequency distributions, interactional contingencies, or socially stratified usage patterns, explanation must extend beyond abstract structural possibilities to include the observed preferences and constraints attested in actual language use. In this sense, naturally occurring data function as a testing ground for typological hypotheses, allowing researchers to assess the robustness of generalizations across registers, interactional settings, and speaker populations.

Typology thus increasingly operates at the interface between structural comparison and empirical modeling of linguistic behavior, as reflected in recent work on distributional and token-based typology (Bickel 2015; Levshina 2019). Far from weakening typological explanation, this shift aligns explanatory claims more closely with observable linguistic behavior and enhances their empirical accountability.

Discourse, interaction, and the emergence of structure. A further cluster of contributions foregrounds discourse and interaction as central sites for the emergence and stabilization of typologically relevant structure. Research in interactional linguistics and conversation analysis has long emphasized that grammar is not merely reflected in discourse, but is partly constituted through recurrent sequential environments and turn design (Hopper 1987; Schegloff 2007; Couper-Kuhlen & Selting 2018).

This perspective has several implications for typological modeling. First, it challenges sentence-centered assumptions in domains where the relevant unit of organization may instead be the turn, the turn-constructional unit, or multi-unit sequences. Second, it foregrounds prosody and temporality as integral components of grammatical organization rather than extragrammatical phenomena. Third, it provides access to intermediate and emergent structures that are crucial for understanding grammaticalization pathways and structural diversification.

Discourse-sensitive analyses thus expand the empirical reach of typology by revealing strategies and constraints that remain opaque in elicited data or exclusively written sources. In doing so, they contribute to a deeper understanding of how typologically comparable structures develop, stabilize, and diverge.

Variation, social anchoring, and typological structure. A point of convergence across the contributions concerns the integration of variation and social structure into typological inquiry. Typological datasets have often treated languages as internally uniform systems; yet decades of variationist sociolinguistics have demonstrated that linguistic structure is systematically shaped by speaker communities, styles, registers, and social meanings (Labov 1972; Eckert 2008). If typology aims to characterize what a language is like, it must also ask for whom, in which settings, and under which communicative pressures.

From a typological perspective, this entails recognizing that variation is a potential source of explanatory insight, particularly when modeled quantitatively across socially defined varieties (Wälchli 2009; Trudgill 2011). On the one hand, it can reveal competing constraints and latent options within grammatical systems, on the other hand socially conditioned distributions shed light on processes of diffusion, stabilization, and restriction. Treating variation as structured evidence thus broadens typology's explanatory scope and improves the comparability of findings across languages, corpora, and communities.

Linguistic Typology at the Crossroads as an editorial space. The orientation emerging from this issue is closely aligned with the editorial positioning of *Linguistic Typology at the Crossroads*. From its inception, the journal has aimed to function as a venue for dialogue rather than as a platform tied to a single theoretical framework. The notion of *crossroads* is therefore not understood as a transitional metaphor, but

as an analytical space in which different approaches to linguistic comparison can meet under shared standards of empirical and theoretical explicitness.

Within this editorial vision, typology is conceived as a field that benefits from methodological plurality, provided that such plurality is accompanied by careful reflection on data, categories, and inferential practices (Schnell & Schiborr 2022). The present issue exemplifies this orientation, bringing together heterogeneous contributions that share the aim of systematically accounting for intralinguistic variation within typological research, and conversely of integrating typological perspectives into the study of variation.

From a methodological standpoint, the contribution by Di Garbo et al. highlights the need for a greater social anchoring of linguistic typology, while Mattiola proposes a method arguing for a discourse-sensitive typology. Evans et al. present a typologically grounded case study based on corpus data, whereas Mithun, and Maschler & Polak-Yitzhaki investigate spoken-language phenomena in a way that enables typological comparison.

Other contributions adopt a clearly variationist perspective –such as those by Tagliamonte, Vietti & Cerruti, and Digesto– while employing analytical tools that facilitate dialogue with typological approaches. Finally, Ballarè et al., using a methodology typical of language variation studies, analyze a phenomenon attested in spoken data from a distinctly typological perspective, while Van Hoey et al. challenge some assumptions on synonymy and linguistic complexity through corpus-based data.

References

Ballarè, Silvia & Inglese, Guglielmo (eds.). 2023. *Sociolinguistic and Typological Perspectives on Language Variation.* Berlin & Boston: De Gruyter.

Biber, Douglas. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

Bickel, Balthasar. 2007. Typology in the 21st century: Major current developments. *Linguistic Typology* 11(1). 239–251.

Bickel, Balthasar. 2015. Distributional typology: Statistical inquiries into the dynamics of linguistic diversity. In Bernd Heine & Heiko Narrog (eds.), *The Oxford handbook of linguistic analysis*, 2nd edn., 901–923. Oxford: Oxford University Press.

Bresnan, Joan & Anna Cueni & Tatiana Nikitina & Harald Baayen. 2007. Predicting the dative alternation. In Gerlof Bouma & Irene Krämer & Joost Zwarts (eds.),

Cognitive foundations of interpretation, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Bybee, Joan. 2010. *Language, usage and cognition*. Cambridge: Cambridge University Press.

Couper-Kuhlen, Elizabeth & Margret Selting (eds.). 2018. *Interactional linguistics: Studying language in social interaction*. Cambridge: Cambridge University Press.

Croft, William. 2001. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford: Oxford University Press.

Croft, William. 2022. *Morphosyntax: Constructions of the world's languages*. Cambridge: Cambridge University Press.

Du Bois, John W. 1985. Competing motivations. In John Haiman (ed.), *Iconicity in syntax*, 343–365. Amsterdam & Philadelphia: John Benjamins.

Eckert, Penelope. 2008. Variation and the indexical field. *Journal of Sociolinguistics* 12(4). 453–476.

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language* 86(3). 663–687.

Haspelmath, Martin. 2018. How comparative concepts and descriptive linguistic categories are different. In Daniël Van Olmen, Tanja Mortelmans & Frank Brisard (eds.), *Aspects of linguistic variation*, 83–113. Berlin & Boston: De Gruyter.

Hawkins, John A. 2004. *Efficiency and complexity in grammars*. Oxford: Oxford University Press.

Hopper, Paul J. 1987. Emergent grammar. *Berkeley Linguistics Society* 13. 139–157.

Kortmann, Bernd (ed.). 2003. *Dialectology meets tyoology: Dialect Grammar from a Cross-Linguistic Perspective*. Berlin & New York: De Gruyter.

Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Levinson, Stephen C. 2016. Turn-taking in human communication—origins and implications for language processing. *Trends in Cognitive Sciences* 20(1). 6–14.

Levshina, Natalia. 2019. Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology* 23(3). 533–572.

Levshina, Natalia. 2022. *Communicative efficiency: Language structure and use*. Cambridge: Cambridge University Press.

Schegloff, Emanuel A. 2007. *Sequence organization in interaction: A primer in conversation analysis*. Cambridge: Cambridge University Press.

Schnell, Stefan & Nils Norman Schiborr. 2022. Evaluating quantitative typological methods: Moving beyond impressions. *Linguistic Typology* 26(2). 245–282.

Sinnemäki, Kaius. 2020. Linguistic system and sociolinguistic environment as competing factors in linguistic variation: A typological approach. *Journal of Historical Sociolinguistics* 6(2). 20191010.

Tagliamonte, Sali. 2006. Analysing sociolinguistic variation. Cambridge: Cambridge University Press.

Trudgill, Peter. 2011. *Sociolinguistic typology: Social determinants of linguistic complexity*. Oxford: Oxford University Press.

Wälchli, Bernhard. 2009. Data reduction typology and the bimodal distribution. *Linguistic Typology* 13(1). 77–94.

CONTACT

silvia.ballare@unibo.it
simone.mattiola@unipv.it
caterina.mauri@unibo.it

Linguistic Typology at the Crossroads 5-2 (2025): 1-35 ISSN 2785-0943

Towards greater social anchoring in language typology

FRANCESCA DI GARBO[1], KAIUS SINNEMÄKI[2] , ERI KASHIMA[2,3]
[1]AIX-MARSEILLE UNIV.,CNRS LPL, [2]UNIVERSITY OF HELSINKI, [3]AUSTRALIAN NATIONAL UNIVERSITY

Submitted: 23/05/2024 Revised version: 17/09/2025
Accepted: 30/10/2025 Published: 25/02/2026

Abstract

In this paper we make the case for the further social anchoring of linguistic typology by illustrating recent methodological developments in the field of comparative research on language contact. We begin by discussing similarities and differences between sociolinguistics and language typology and focusing on the issue of the social anchoring of research on language variation and linguistic diversity. We argue that while sociolinguistics is socially anchored by definition, linguistic typology has so far abstracted languages from their social contexts due to the nature of macro-comparison and the poor availability of data on sociolinguistic environments. We suggest that greater social anchoring in large-scale comparative research on language structures is possible and can be achieved through integrating two general aspects of language use, relationality and context-dependency, into typological models of language variation and change. We illustrate how data collection on bilingual language ecologies embodies the notion of relationality and context-dependency.

Keywords: Sociolinguistics; language typology; language contact; language ecology.

1. Sociolinguistics, language typology and the social anchoring of language variation

Sociolinguistics and language typology are among the subfields of linguistics that share an interest in linguistic variation. Despite this shared interest, ever since they emerged as established domains of research in the 1960s (through the work of

William Labov for sociolinguistics and Joseph Greenberg for typology), the methodologies and research materials of these two fields have been quite different.[1]

A key idea in sociolinguistics is that language is a social institution and thus intertwined with different social phenomena. Much evidence has been accumulated showing how language use varies, for instance, across age and gender, how that variation may be harnessed for indexing social identity (e.g., Silverstein 2003), and how socially structured variation may lead to language change over time (e.g., Milroy & Milroy 1985; Croft 2001; Nevalainen & Raumolin-Brunberg 2017). Data is typically collected from individuals, analyzing their language production as well as some relevant social aspects hypothesized to structure production, such as age or gender and, more recently, community of practice (vid., Wenger 1998; King 2019). Macro-level analyses are relatively uncommon, although they appear to be somewhat more common in research on the sociology of language, which can be broadly defined as the study of the relationship between language and society with a focus on the societal aspects of linguistic behavior (Chen 1997).[2]

In language typology, an overarching aim is to better understand the world's linguistic diversity; that is, why the thousands of human languages spoken or signed across the globe are the way they are, what unites them, and how they differ from one another. Research in this field typically draws data from descriptive research materials, such as reference grammars. The focus is usually on different qualitative and quantitative aspects of linguistic patterns rather than on their behavioral profiles. Recently, however, corpus-based typological research that enables measuring usage frequency and assessing behavioral properties has been growing (e.g., Levshina 2019) thanks to the increased availability of multilingual annotated corpora, such as Universal Dependencies (Zeman et al. 2023), MultiCAST (Haig & Schnell 2023) and DoReCo (Seifart et al. 2024). Data from individual language users is rarely collected in typological research, although incipient examples do exist (e.g., Dingemanse et al. 2013).

[1] Francesca Di Garbo and Kaius Sinnemäki contributed to all aspects of the work, from the study design to the write-up. Eri Kashima contributed to the write-up of the work and the design of section 3.2.

[2] Sometimes sociology of language and sociolinguistics are treated as two separate fields despite their considerable overlap. Here we subsume sociology of language under sociolinguistics (following e.g., Bell 2013) but refer to the former where necessary for the sake of clarity.

These two research fields thus differ in various ways. One difference is that of scale. Sociolinguistics tends to focus on micro-level analyses that investigate variation at the level of individual members of a community (population of individuals). Language typology, on the contrary, tends to focus on macro-level analyses that study variation across languages (population of languages). In other words, while, in sociolinguistics, the units of comparison tend to be individuals and their sociological and linguistic profiles, in language typology, the units of comparison are languages or specific constructions in given languages, essentially viewed as shared communication systems.

Another major difference between these fields is in the social anchoring of research on language variation. What we mean by anchoring is whether research on language variation is informed by what is known about the social contexts in which variation is embedded; that is whether our models of variation and change do justice to the social nature of the linguistic phenomena observed.[3] Languages do not exist in a vacuum but in different sociocultural and geopolitical contexts that are in turn anchored in human behavior. Ideally, our models of language variation should account for those contexts to evaluate the extent to which human behavior, including linguistic behavior, may depend on those contexts. Because of its commitment to the individual language user and the social aspects of language use, social anchoring is a theoretical as well as a practical aim in sociolinguistic research. However, research in language typology tends to lack this kind of anchoring. Typological models are usually based only on language descriptions without a connection to the social contexts in which the described patterns occur.

[3] The notion of *anchoring*, as we use it in this paper, may somewhat overlap with that of *grounding*, as used in philosophy. In philosophy, the notion of grounding is understood as a metaphysical, noncausal relation between some aspect of reality that gives rise to or is the basis for some other aspect of reality: when A is grounded in B, A depends at least partially on B (see e.g., Correia & Schnieder 2012). Thus, saying that language is grounded in the social reality of speakers and signers is making a claim that some properties of language depend, at least to an extent, on the language users, albeit not causally but constitutively. While this may be a reasonable position when approaching some aspects of language, our interest in the relationship between language variation and the social context of language use concerns the probabilistic and possibly causal relationship between the two. We are specifically interested in whether language structures may adapt to the sociolinguistic context in which languages are learned and used (see Di Garbo et al. 2021; also Adamou 2021). We decided to use the more theoretically neuter term *anchoring* in order to avoid any potential pitfall related to using the more connotated term *grounding*.

Only in the past two decades have societal factors started to play a role in a research area now called sociolinguistic typology (e.g., Trudgill 2011; see the review in Sinnemäki, in revision). This research may integrate aspects of the language community into their models, such as size of the language community or intensity of language contact. But, otherwise, factors related to language users tend to play no direct role in data collection and analysis in language typology. Some such factors, such as cognitive processing preferences and learning constraints, may be used for explaining typological distributions but typically only in a post hoc fashion, rather than by being built into the framework (e.g., Lupyan & Dale 2010; but see research on efficiency, for instance Levshina 2022, where cognitive preferences may be used to at least generate hypotheses that are then tested with typological data). Furthermore, these factors seem to be limited mostly to individual cognition, such as memory limitations, rather than to social aspects of language use (see Kusters 2003 for an early example of this approach, focusing on second language learning and linguistic complexity in the verbal domain).

There are both historical and methodological reasons for this situation in language typology. As outlined by Bickel (2007), up until the late 20th century, typological research often focused on the question of what is possible in language. The aim was to find universal cognitive constraints, and typology was often seen as the flipside of research on universal grammar. Language was largely abstracted away from its social environment and attempts to correlate language structure and language-external factors were approached with strong reservations (see e.g., the discussion and references in Ladd et al. 2015: 227). Since roughly the 1990s there has been a steady movement into appreciating variation at all levels, giving typology a much more anthropological orientation (Bickel 2007). As a result, research in modern language typology tends to be probabilistic and embraces the fact that most typological features are geographically and genealogically distributed. Language universals are seen as structural pressures that affect how languages change over time rather than as cognitive constraints (e.g., Greenberg 1978; Sinnemäki 2010; Dunn et al. 2011; Bickel 2013). This “anthropological” shift with an appreciation of local variation has been a boon that has advanced research in the field (the journal *Linguistic Typology at the Crossroads*, with its programmatic effort to encourage debate around typology and its bordering disciplines, is a good example of this process). Along these lines many typologists also assume that the language system is not autonomous from language users but embedded in and affected by various language-external factors, such as

language use, social interaction and cognitive processing (e.g. Trudgill 2011; Hawkins 2014; Sinnemäki 2014; Pakendorf et al. 2021; Levshina 2022).

Despite this theoretical stance, linguistic patterns tend to be analyzed in typological research as if they were autonomous of language users. As mentioned above, language users enter the research design only post hoc, if they enter at all. Crudely said, and we are being self-critical here as well, especially in the field of large-scale quantitative typology, the data type that may be the most "social" seems to be the sheer geographical location of languages in terms of longitudes and latitudes, often represented as focal points. Geographical locations may then be used as measures of geographic distance or as a proxy for evaluating or controlling for the effects of language contact (see, for instance, Guzmán Naranjo & Becker 2022; Guzmán Naranjo & Jäger 2023). But language contact is founded on bilingual and multilingual interactions that are embedded in the social fabric of language use. It is also not clear how the geographical distance between language communities may be related to the social environment of language use more broadly, and this could in fact be turned into a research question on its own.

In this article, we build on the ongoing anthropological shift in language typology and argue that there are good theoretical grounds for taking this shift even deeper through a better social anchoring of research in the field. This social anchoring would involve, for instance, better addressing typical concerns in the sociology of language such as "who speaks what language to whom and when" (Fishman 1965), and thus producing descriptive data on the social environment of language use. In the next sections, we argue that such a shift in language typology is methodologically feasible and illustrate how it can be implemented concretely. We tackle the relationship between language structures and language use patterns by focusing on language contact dynamics and the impact of multilingual language ecologies on patterns of language structures. In language contact studies, there is a long tradition of research focusing on explaining the scenarios underlying the outcomes of language contact (see, for instance, Muysken 2013; Ross 2013). Thus, language contact phenomena provide a promising laboratory for investigating the social anchoring of linguistic diversity.

The paper is structured as follows. In Section 2, we discuss in more principled terms the conditions for increasing the social anchoring of typological research to base it more on naturalistic data. Two principles are introduced to this effect, what we call *relationality* and *context-dependency*. In Section 3, we demonstrate how we implement

these principles in a research design that is geared to capture contact-induced language change in bilingual language ecologies across the world. Section 4 ends the article with discussion and brief conclusions.

2. How to increase the social anchoring of research in language typology

2.1. On abstraction in language typology

In any field of research, comparison implies some level of abstraction. This is necessary in order to navigate individual manifestations of a given phenomenon while, at the same time, searching for patterns in the distribution of similarities and differences between data points (for an overview of the epistemology and history of the comparative method in the Humanities and Social Sciences, see Griffiths 2017).

In comparative linguistics, and language typology in particular, abstraction stems from at least two facts. Firstly, in the large majority of cases, the focus of comparison is not on languages as manifested through the repertoires of individual users, but rather on descriptive resources. These provide a reified representation of the functioning of said languages, what typologists refer to as *doculects* (Cysouw & Good 2013). Secondly, and related to the first point, given that the focus of comparison is on constructions as represented and described in a doculect, the generalizations stemming from large-scale typological research are inevitably limited to the linguistic properties of these constructions rather than encompassing the actual usage preferences of the population of individuals who associate themselves with a given language.

In sum, the population that is being compared in typological studies comprises a selection of doculects or, more specifically, a selection of constructions in a sample of doculects. Thus, even though, broadly speaking, many typologists conceptualize languages as communicative systems embedded in their own socio-historical, cultural and environmental ecologies, large-scale typological research tends to have a much broader and somewhat abstract focus, given that the object of study, languages and their constructions, are approximated through their documentary description, that is, doculects. Given this background, and compared to other fields of linguistic research, such as pragmatics, conversational analysis and sociolinguistics, the social agency of speakers probably cannot be operationalized in typology at the same level of detail. Typological data alone have thus far offered simplified representations of language

variation because they tend to leave out the social foundations of variation at the level of the individual.

However, this abstracting away from the social foundations of variation may also depend on the state of the art in data availability in language typology. Typology is about comparison, but comparison is possible only when sufficient and comparable data is available for doing so. Linguistic descriptions have been available for several hundred languages already for decades, making it possible to collect large typological datasets on linguistic features. However, descriptive and comparative work on the sociolinguistic environments of language use has been much slower and, thus far, it has also prevented language typology from greater social anchoring (Sinnemäki, in revision), However, the fact that, for instance, recent descriptive grammars sometimes provide actual data on the social correlates of structural variation at the community level, as in the work by Kluge (2017) on Papuan Malay (pmy; Austronesian, Nuclear Malayic), testifies of a growing interest in the social anchoring of language use.

Yet another sign of this ongoing paradigmatic shift towards a better integration between, on the one hand, language variation and change at the community level and, on the other hand, worldwide linguistic diversity is the work by Mansfield et al. (2023) on grammatical variation between closely related dialects in a typological perspective. They investigate how the social-signaling role of grammatical variation may contribute to linguistic divergence and diversification. This is done using a sample of 42 languages across the world and based on documented descriptions of dialectal variation in these languages. The study finds that dialect differentiation between closely related varieties in close contact is mostly carried out by what they call *form-variables*. This is when dialectal difference is manifested by variant forms of the same morpheme across dialects, such as the negative auxiliary/copula variation in spoken varieties of English[4] (*It **isn't** right* vs. *it **ain't** right*). This variation is interpreted as a sign of the strong social indexing role that linguistic divergence plays in situations of dialect contact. The method developed by Mansfield et al. (2023) represents a groundbreaking innovation as it shows that dialectal differences, which are usually investigated in variationist sociolinguistic and dialectological studies of individual speech communities, can be also studied from a large-scale typological perspective.

In this paper, we argue that increasing the social anchoring of large-scale typological investigations is also possible when comparing genealogically unrelated languages. However, we argue that this step requires some further level of

[4] eng; Indo-European, Germanic.

abstraction. We therefore propose a new type of heuristics in the study of linguistic diversity and one which is anchored to the relational and context-dependent nature of linguistic phenomena, as presented in the next subsection.

2.2 The relational and context-dependent aspects of language use

We propose two general principles related to the social aspects of language use and argue that they both provide a viable starting point for increasing social anchoring in language typology. We also demonstrate how these principles could be built into a typological approach in practice. These two principles are (1) the relational and (2) the context-dependent nature of language use. First, humans have a disposition to interact and cooperate with one another, to share intentions, and to form coordinated group activities (e.g., Tomasello 2010; among many others). All these aspects are closely tied to language and its use: one of the main functions of language is communication, which by definition is relational, interactional and cooperative. Second, language use is strongly context-dependent, which means that human communicative practices vary depending on the situation, on the audience, and even on eavesdroppers. This context-dependency is the basis for much linguistic variation, such as different styles, genres and registers (e.g., Biber 1988; among many others).

In our typological research, we have turned these two general social aspects of language use into two interrelated research foci about language ecology, as in (1). The term language ecology here broadly refers to the interaction between language and its environment (Haugen 1972), but we limit our discussion to the social and sociohistorical aspects of the environment.

(1) Our operationalization of language ecology
a. Relationality → Focus on bilingualism and language contact
b. Context-dependency → Focus on variation in language ecology and in language use

In terms of the relational aspect of language use, we focus on bilingualism and language contact, that is, how language changes when interaction takes place between people speaking different languages. We approach the context-dependent aspects of language contact in the following two ways, both focusing on variation. First, we research the social aspects of the bilingual language ecology in different

social domains, such as the family or the occupational context, and not simply overall in the community (e.g., Fishman 1965; Di Garbo et al. 2021). Second, we attempt to capture linguistic diversity in the spirit of multivariate typology, which proposes fine-grained typological variable design as a way of integrating information about language-internal variation in large-scale comparative research on the world's languages (e.g., Witzlack-Makarevich et al. 2022). Through this demonstration, we aim at illustrating what contribution this newly developed approach brings to the typologist's toolkit, by making a step forward towards a principled understanding of linguistic structures as embedded in their language ecologies.

Applying these two principles concretely would mean creating datasets that incorporate as much information as possible about language-internal variability in the distribution of linguistic features. It also means factoring in characterizations of the sociolinguistic environments of language communities. Capturing language-internal variability is a way of getting at the relational aspect of language use, that is to the fact that linguistic variants are constantly negotiated through interactions between individuals and the representations that people build of their interlocutors (relationality). In our work, we specifically focus on developing and testing methods that can allow us to detect how bilingual language use affects the emergence and development of linguistic variants. For instance, if several alternative strategies of encoding are attested for one and the same linguistic feature, could any of these have emerged as a result of contact with neighboring communities? At the same time, characterizing the social embeddings of linguistic interactions provides a way of assessing how the distribution of linguistic variants may also depend on the specific socio-cultural contexts in which these interactions are situated (context-dependency). For instance, can language dominance or language ideologies tell us anything about the type of linguistic variants that are more likely to emerge in contact situations?

The relational and context-dependent principles of language use have been the objects of recent discussions in small-scale multilingualism research (for an overview see Pakendorf et al. 2021). Linguistic identities in small-scale multilingual societies are characterized as multifaceted and sensitive to contexts of interaction. These contexts of interaction are highly localized and can be based on factors such as place (e.g. for speaking the languages associated with specific places see Merlan 1981 for Australia and Döhler 2018 for southern New Guinea) or relationship with interactant (e.g. for speaking to in-laws, see Fleming 2011 and for speaking to clan members, see

Garde 2008; Suokhrie 2016; for speaking a village language of communication, see Gumperz & Wilson 1971).

Additionally in those small-scale multilingual societies where the linguistic repertoire of language users coincides with a pool of closely related varieties, it has been observed that minimal structural differences across varieties may often convey strong social meanings, associated with one or the other speaker community. Lüpke (2022) illustrates these processes of socially charged linguistic divergence in the nominal classification systems of the languages of Lower Casamance (Senegal). These closely related varieties typically display the same inventories of noun class distinctions from a formal point of view. However, the assignment of individual nouns to such classes differs across varieties and these differences actually index people's 'belonging' or 'identifying' themselves as members of one or the other community. Examples are shown in Table 1.

Baïnounk Gubëeher	**Joola Kujireray**	**Meaning**
bu-óóg/i-óóg	*fu-bah/ku-bah*	'baobab fruit(s)'
bu-gof/i-gof	*fu-how/ku-how*	'head(s)'
bu-koor /-i-koor	*e-suh/si-suh*	'village(s)'
bu-deen	*e-baŋ*	'putting'

Table 1: Mismatches in noun class assignment across two closely related varieties of Lower Casamance (Senegal). Examples taken from Lüpke (2022)

In Baïnounk Gubhëeher (gube1234; Altlantic-Congo, North-Central Atlantic)[5] and Joola Kujireray (bkj; Altlantic-Congo, North-Central Atlantic), two languages spoken in Lower Casamance, the nouns class markers *bu-/i-* (Baïnounk Gubëeher) and *fu-/ku-* (Joola Kujireray), respectively, are semantically equivalent cognate classes that are used for the classification of round things. As suggested by the examples in Table 1, the same noun class markers are used in these languages for some words, such as the nouns for 'baobab fruit(s)' and 'head(s)'. However, nouns that are assigned to this semantically motivated class in one of these varieties may end up being assigned to other classes in the other, such as the nouns for 'village', which belongs to class *bu-*

[5] This language does not have an ISO-code; the Glottocode is used instead.

/i- in Baïnounk Gubëeher but is assigned to class *e-/si-* in Joola Kujireray. Such differences in class assignment across closely related varieties may then be used for indexing people's 'belonging' to one or the other community. For analogous examples, in a different small-scale multilingual setting within Western Africa, the Cameroonian Grassfields, see also the recent study by Di Carlo & Good (2023).

Our claim is that the relational and context-dependent aspects of language use, which have already been identified as relevant for understanding and modelling types of multilingualism, are widely applicable to any situation of contact between users of different languages. Accounting for these aspects of language use in typological studies could thus provide a starting point for increasing the social anchoring of the crosslinguistic generalizations that are made in these studies.

Once this possibility is acknowledged, the onus is to make explicit and workable steps towards turning these observations into implementable methodologies. In particular, in order to turn this fluid and versatile representation of the workings of language into something that can be compared across time and space, some form of reification becomes necessary. For instance, while in linguistic research there is a growing call to shifting the focus from *languages* as compartmentalized entities to *languaging* as the constant negotiation of communicative repertoires in context (see, e.g., Lüpke 2024), comparing linguistic structures, as used by a population, still requires some sort of schematization of the phenomena being compared, be it a set of specific constructions or a holistic representation of a linguistic code. Importantly, psycholinguistic studies suggest that reified representations of languages and linguistic practices also have some kind of psychological reality (see, for instance, Berthele 2021). That is, a linguistic code may be a clearly identifiable object for multilingual speakers, and the separability of the code(s) may be what enables languaging in the first place.

In section 3 we illustrate how the relational and context-dependent nature of language use may be operationalized typologically in the crosslinguistic study of language contact.

3. Language contact in its bilingual language ecology: some illustrations

In the previous sections we argued for the theoretical importance of increasing social anchoring in language typology. We discussed the methodological requirements for doing so and the reasons related to the lack of such anchoring in earlier typological research. We suggested that the relational and context-dependent nature of language

could be important starting points to address when linking language typology to the social aspects of language ecology.

Here we will illustrate how these principles have been built into a new typological research design stemming from research conducted during the ERC Starting Grant project GramAdapt (e.g., Di Garbo et al. 2021). While this illustration relies on research conducted or published in the context of this larger project, the discussion of the principles of relationality and context-dependency is original to the present paper. We discuss how each of the two principles has been built into the broader research design of the project, and how ensuing methodological issues have been addressed.[6]

3.1. Relationality and context-dependency in GramAdapt linguistic data

To research the relational nature of language use, a suitable linguistic phenomenon is needed that could be feasibly compared across languages. We argue that language contact and bilingualism offer one such area. Contact and bilingualism are relational from the outset, because in bilingual ecologies two or several populations of individuals speaking different languages may interact with one another and may thus also influence one another's linguistic behavior.

To assess such influences across languages, we have developed a new typological approach to language sampling, which is geared to make inferences about contact-induced change (see Di Garbo & Napoleão de Souza 2023 for details). Languages are selected in pairs based on prior evidence of interaction between the language communities of interest. The primary language of interest in any given pair is identified as the Focus Language, and its contact language is the Neighbor language. The pairs of Focus and Neighbor Languages form a "test case" that lets one zoom in on the relational nature of contact. An example of a test case is the contact between Alorese (aol; Austronesian, Bima-Lembata) and Adang (adn; Timor-Alor-Pantar, Nuclear Alor-Pantar) in the East Nusantara region of Eastern Indonesia (see Figure 1). Notice that Alorese is in contact with other Alor-Pantar languages spoken in the Pantar Islands and its islets. Here we follow the work by Moro (2021) who focuses on

[6] There are also other approaches to implementing the relational nature of language use into typological analysis, for instance, in pragmatic typology (see e.g., Floyd et al. 2020 and Rossi et al. 2020). However, those approaches focus on interactional dynamics and thus assume finer-grained data than what is possible in the context of the approach presented here, which is geared towards studying structural features of languages.

the Alorese-Adang contact scenario because of its neater dynamics in comparison to the other neighboring languages. We refer to Moro's paper for an in-depth study of the multilingual patterns of the Alorese community.

To analyze whether contact with the Neighbor language has led to changes in the Focus language, a "control case" is also selected; a language that is closely related to the Focus language but that has not been in contact with either it or the Neighbor language (Di Garbo & Napoleão de Souza 2023). As an example, the western varieties of the Austronesian language Lamaholot (slp; Austronesian, Bima-Lembata) would serve as a reasonable control case, since it is known that Alorese split off from western Lamaholot no later than roughly 600 years ago (Klamer 2012). At best, such a control case, or Benchmark, may approximate the state in which the Focus language was prior to its contact with the Neighbour (e.g., Sinnemäki & Ahola 2023; Sinnemäki et al. 2024). In this paper, we choose the Leiwong dialect of Lamahalot, described by Nishiyama & Kelen (2007), as a Benchmark.

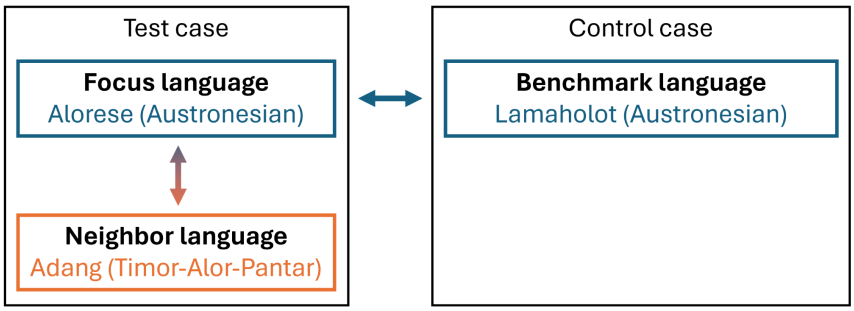


Figure 1: "Test case" and the "control case" according to the sampling scheme of Di Garbo & Napoleão de Souza 2023).

This approach to language sampling shifts the attention from viewing languages in isolation to incorporating the relational nature of language contact into the model. The triplet Focus-Neighbor-Benchmark also enables us to draw dynamic inferences about the outcomes of language contact (for earlier studies in dynamic typology see, e.g., Maslova 2000; Maslova & Nikitina 2007; Cysouw 2011; Bickel 2013). Since more than one language is selected from the same family, it is possible to draw inferences about type shifts, that is, how the Focus language may have changed in the contact situation, and as compared to the Benchmark sister language.

Focusing on language contact from a worldwide comparative perspective may raise some methodological issues. For instance, a basic requirement in quantitative

research is that datapoints are independent of one another (see Winter & Grice 2021 for a recent discussion). Yet, when languages are sampled in contact pairs, they cannot be considered fully independent of one another. The Focus and the Benchmark also come from the same language family and therefore are not independent of one another. This point is broadly shared by all diachronically minded approaches to typological research, such as dynamic typology (e.g., Bickel 2013): when typological universals are interpreted as diachronic pressures to type shift, in order to assess the dynamics of these type shifts, it becomes necessary to select two or more languages from the same family. Classical statistical tests therefore cannot be applied to the data sampled in such a way.

In our approach, the datapoints are not linguistic features of individual languages but potential contact-induced type shifts (and "non-shifts", see below) in the Focus Language which are identified and analyzed by taking into account what is attested in the respective Focus-Neighbor-Benchmark triplet. Our approach is thus inherently dynamic in that it understands language structures as constantly evolving. In order to capture these patterns of evolution, the linguistic variables of interest are situated in a broader context by taking into account what they could have become if contact-induced change occurred (i.e. given structure as realized in the Neighbor language), and by comparing them with analogous structures as attested in a close relative (the Benchmark).

As an illustration, let us consider the order of the possessor and possessum in alienable possessive constructions in our example set (Alorese-Adang-Lamaholot). In the Benchmark Language Lamaholot, the possessor follows the possessum, as in (1). In the Neighbour Language Adang, the possessor precedes the possessum, as in (2). The Focus Language Alorese has developed the possessor-possessum order, as in (3), possibly via contact with Adang. It is thus the type shift from the possessum-possessor order to the possessor-possessum order in Alorese that is the independent data point to count in these constructions.

(1) Benchmark: Lamaholot (Bima-Lembata, Austronesian; Nishiyama & Kelen 2007: 23)
lango *go'en*
house 1SG.POSS
'my house'

(2) Neighbour: Adang (Timor-Alor-Pantar; Haan 2001: 147)
n-ɔ *baŋ*
1SG-GEN house
'my house'

(3) Focus: Alorese (Bima-Lembata, Austronesian; Sulistyono 2022: 106)
go *uma*
1SG house
'my house'

A non-shift, in turn, is simply a situation when the Focus Language has retained the allegedly inherited feature. The feature is also attested in the Benchmark Language but not present in the Neighbor language. For instance, the nominative/accusative form of the first-person pronoun in the Benchmark Lamaholot is *go*, while its genitive form is *go'en* (Nishiyama & Kelen 2007: 13). The first-person genitive pronoun in the Neighbour Language Adang is *nɔ* (or *ne*; Haan 2001: 149). In the Focus Language Alorese, the first-person form of the pronoun is *go* and this form is used for subject, object and possessive functions. In other words, Alorese seems to have retained the form of the first-person singular pronoun *go*, although it has also lost its genitive form.

As long as the triplets are independent of one another, it makes sense to assume that the distribution of type preferences gleaned from the triplets are also independent of one another. Thus, drawing inferences from the triplets does not preclude the use of statistical methods. Even if one was unwilling to use classical statistical methods for such data, logical inferences can still be made about such distributions with the help of Monte Carlo methods (e.g., Janssen et al. 2006).

Integrating the relationality principle to sampling has thus significant repercussions to data collection: to make inferences about language contact, linguistic data need to be collected from each contact pair but also from each Benchmark Language. This data collection can be done through standard practices in language typology; for instance, by analyzing descriptive linguistic data collected from reference grammars.

In addition to factoring in the relational aspect of language use, by looking at pairs of languages in contact, our method also lets us explore contextual variability in linguistic structures. We achieve this by working with a coding design that is purposely geared towards exploring patterns of language internal variation for any

given linguistic variable (Sinnemäki et al. 2024). This coding design is inspired by the principles of multivariate typology (Witzlack-Makarevich et al. 2022), whereby linguistic structures are explored through fine-grained questions that contribute to depicting and characterizing variation. For instance, with respect to nominal number, we do not just look at whether a given language marks the plural, but whether the plural is marked on nouns, pronouns of various types, adnominal modifiers, verbs etc. And, with respect to nominal number marking, we do not just ask whether this is suffixal or prefixal but whether different types of markers (suffixes, prefixes, stem alternations, reduplication) may occur on nouns or through agreement (Di Garbo & Kapellis 2025). Through this procedure, we can detect whether languages exhibit different strategies in different domains and whether this speaks of any ongoing change. We indeed find internal variation in the distribution of linguistic features of the Focus languages. When comparing this variation with the constructions attested in the respective Neighbor and Benchmark languages, these patterns of internal variation can, in some cases, be interpreted in terms of contact-induced type shifts.

Coming back to the Alorese-Adang contact pair as an illustration, similarly to its Neighbour Adang, Alorese marks nominal plurality by means of a plural word, *hire*, which is historically derived from third person plural pronouns. This is shown in (4) and (5). Conversely, the Benchmark Lamaholot does not have plural words, but nominal plurality is optionally marked through reduplication, as shown in (6).

(4) The plural word *hire* in Alorese (Moro 2018: 184)

məsia	*hire*	*ke*
person	PL	DEM.PROX

‘these/the persons’

(5) The plural word *nun* in Adang (Haan 2001: 122)

pen	*ti*	*mat*	*nun*	*ʔa-bɔʔɔi*
Pen	tree	big	some/several	3.OBV-cut

‘Pen cut some big trees.’

(6) Plural reduplication in (Lewoingu) Lamaholot (Nishiyama & Kelen 2007: 210)

inamvlake-inamvlake	*svga-ka*	*urin*
man-man	came-3PL	late

‘Men came late.’

A few reduplicated nominal stems also exist in Alorese, but their non-reduplicated bases are no longer attested. This suggests that Alorese also used to mark plurality through reduplication as does its sister language Lamaholot, but this strategy was later replaced by the emergence of plural words, as in the Neighbor language Adang. In this case the inferences drawn from our coding method can be backed up by existing literature. Moro (2018) demonstrates that the grammaticalization of the Alorese plural word is indeed the result of a contact-induced type shift in the wider historical context of contact between Austronesian and Alor-Pantar languages. Additional examples of language internal variation in nominal number systems which are suggestive of ongoing contact-induced type shifts in the languages of our sample can be found in Di Garbo & Kapellis (2025).

These examples illustrate how we address contextual variability in linguistic structures through fine-grained analyses of the structures attested in the Focus languages and then comparing them with data from the respective Neighbor and Benchmark languages. We argue that capturing language-internal variation in this way is one of two possible ways of exploring context-dependence in contact-induced variation from a typological perspective. Another way, which we have not yet implemented in our own work, is to compare language structures in closely related dialects with different contact profiles to test whether dialectal variation can be eventually explained as a function of proximity or degree of contact with a genealogically unrelated neighbor (see Mansfield et al. 2023 for a recent advance in dialect typology, also discussed in section 2.1) These two approaches make the typological study of language variation and change in contact situations more dynamic; that is, by capturing language-internal variation through comparisons of genealogically unrelated contact pairs on the one hand, and by comparing dialect pairs with different contact profiles on the other. As we hope to have shown, the two approaches have the potential to significantly contribute to increasing the ecological validity of typological generalizations.

3.2 Relationality and Context-dependence in GramAdapt sociolinguistic data

In our approach to contact and bilingualism, we also factor in the sociolinguistic aspects of the bilingual language ecology. Collecting such data is, however, hindered by poor data availability. There is little descriptive data available on sociolinguistic environments in general, and on bilingual language ecologies in particular.

Descriptive grammars often contain sections on language ecology, but their extent varies and their usefulness for evaluating bilingual language ecologies in particular may be low. Other initiatives, such as the articles in the Language Context section of the journal *Language Documentation and Description* may contain more useful data but tend to focus on a community's use of a single language. Overall, the descriptive status of sociolinguistic environments of languages is rather low.

For this reason, we developed a sociolinguistic questionnaire within the project, with the aim of eliciting fine-grained descriptions of the Focus-Neighbor contact profile at a particular point in time (see Kashima et al. 2025 for an overview of the questionnaire design). The point in time was defined as when there were the most opportunities for interaction between the Focus and Neighbor language speakers. The sociolinguistic aspects of contact are thus analyzed between selected individual languages within the "test case". The Benchmark Language was selected specifically because it has not been in contact with the Focus or the Neighbor and hence no social contact data was collected on it. The questionnaire was directed to specialists such as documentary linguists and anthropologists who are well familiar with the Focus-Neighbor contact situations. Here, we focus on the methodological choices that we made in order to tie this questionnaire to the relational and context-dependent aspects of language use and also illustrate initial results. For the purpose of this illustration we draw data from Kashima et al. (2023), which is a dataset of 34 responses to the sociolinguistic questionnaire, featuring 34 contact scenarios from around the world and partially overlapping with the sample presented by Di Garbo & Napoleão de Souza (2023). Figure 2 shows the distribution of the contact scenarios.

Contact research was at the heart of the first attempts to increase the social anchoring of typology when sociolinguistics and language typology were first being integrated more than 20 years ago (see Sinnemäki, in revision). The most commonly incorporated sociolinguistic data in these studies were population size and some broad qualitative assessment of the intensity of contact (e.g. Sinnemäki 2009; Lupyan & Dale 2010; Bentz & Winter 2013; Sinnemäki & Di Garbo 2018). Thus, although some aspects of social contact were integrated in typological models, these factors were analyzed on a general level and without investigating contact relations between specific languages, that is, in an essentially non-relational way (see below). For instance, Bentz & Winter (2013) collected data on the proportion of non-native

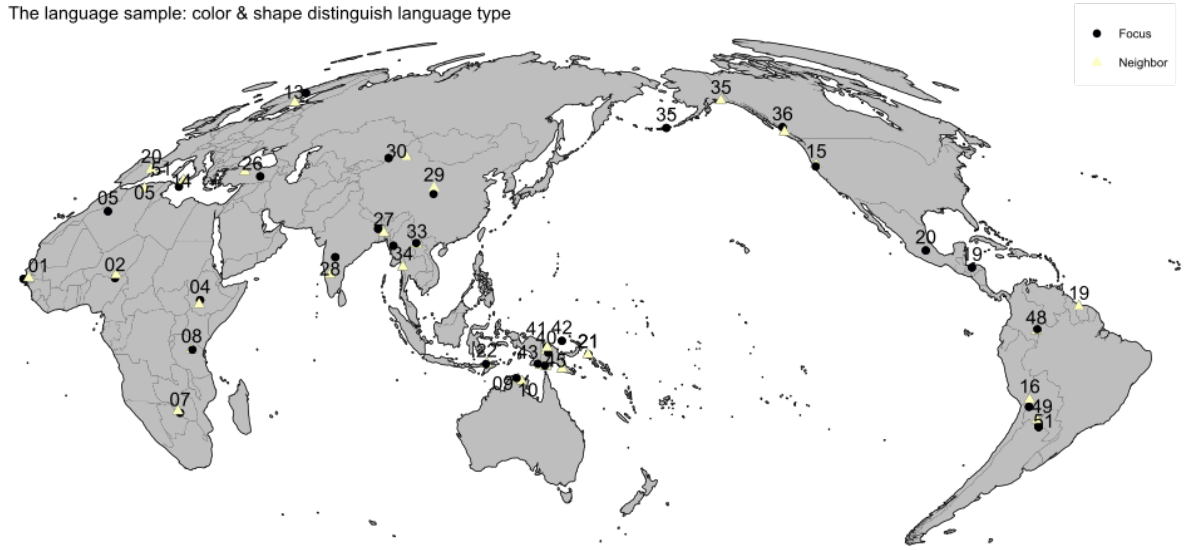


Figure 2: the 34 contact scenarios of the sociolinguistic dataset. Focus languages are represented as black circles and Neighbor languages as beige. Contact pair IDs are also plotted.

speakers in whole speech communities (see also Trudgill 2011). Such general-level demographic data provide information about the overall extent of bilingualism in the community and could potentially be connected to mechanisms of language change that depend on such general-level phenomena. If, for instance, the speaker population is accustomed to many people from different backgrounds learning their language, this could mean that the native population might also be accustomed to producing foreigner-directed speech, which is one of the mechanisms of language change assumed to lead to simplifications (e.g., Kusters 2003; Trudgill 2011; Bentz & Winter 2013; Berdicevskis 2020).

A concrete example of a non-relational way of describing a bilingual ecology could be the phrasing given in (7a) while (7b) illustrates how this example could be phrased in terms of the specific contact situation between Alorese and Adang.

(7) Non-relational ways of describing language contact
a. “Speakers/signers of language X are typically bilingual.”
b. “Speakers of Alorese are typically bilingual.”

This statement could then prompt a binary yes/no response, a set of Likert-scale type predefined values ranging from "very uncommonly" to "very commonly", or perhaps even a numerical value that would represent the extent of bilingualism in the community. Note that such a non-relational variable abstracts away from the specific contact relations between language communities and may result in loss of relevant information about the contact ecology.

In our approach, we build the relational nature of language contact into the model in terms of the sociolinguistic aspects of contact. We thus assume that contact-induced changes may be affected by the particular Focus-Neighbor contact ecology as well as by the properties of the Neighbor language. This emphasis on the relational aspect of the bilingual ecology implies developing ways to tackle the relational nature of a given contact situation, which is often quite straightforward. For instance, a non-relational variable, such as (7a), can be easily turned into a relational one by zooming in on a particular contact scenario, as in (8a). Examples (8b) and (8c) illustrate how these variables could be phrased in terms of our example test case involving Alorese and Adang.

(8) Implementing the relational aspect of language contact
a. "Speakers/signers of language X are typically bilingual in language Y."
b. "Speakers of Alorese are typically bilingual in Adang."
c. "Speakers of Adang are typically bilingual in Alorese."

If both linguistic and sociolinguistic data were collected on language X and Y, this would enable making inferences about how the social contact between the two language communities has potentially affected the languages in that contact scenario (see more in Sinnemäki & Kashima, forthcoming).

We illustrate how the relational aspect of social contact was built into our questionnaire, by drawing one example from Kashima et al. (2023). This example question asks about language attitudes towards linguistic transfer. A non-relational variable probing language attitudes to linguistic transfer could ask what the Focus language speakers' attitude to lexical or grammatical borrowing are on a general level, regardless of the source contact language. However, such attitudes often depend on the language at stake, and, for this reason, it makes sense to ask about attitudes in a relational way, that is, by focusing on specific contact pairs. For example, the notable difference in the prevalence of Swedish (swe; Indo-European, Germanic) vs

Russian (rus; Indo-European, Slavic) origin loanwords in Finnish (fin; Uralic, Finnic) throughout the twentieth century (Cronhamn 2018) is in part explained by the more positive attitude that Finns historically had towards Sweden and Swedish rather than towards Russia and Russian. Question ID OI6 in Kashima et al. (2023) thus asks the question in (9).

(9) Example question from the sociolinguistic questionnaire
"What are the Focus language speakers' attitudes towards linguistic transfers from the Neighbor language, such as lexical or grammatical borrowing?"

The predefined responses to this question are on a Likert scale and range from "Very negative" to "Very positive". The responses for 34 Focus-Neighbor pairs are summarized in Figure 2.

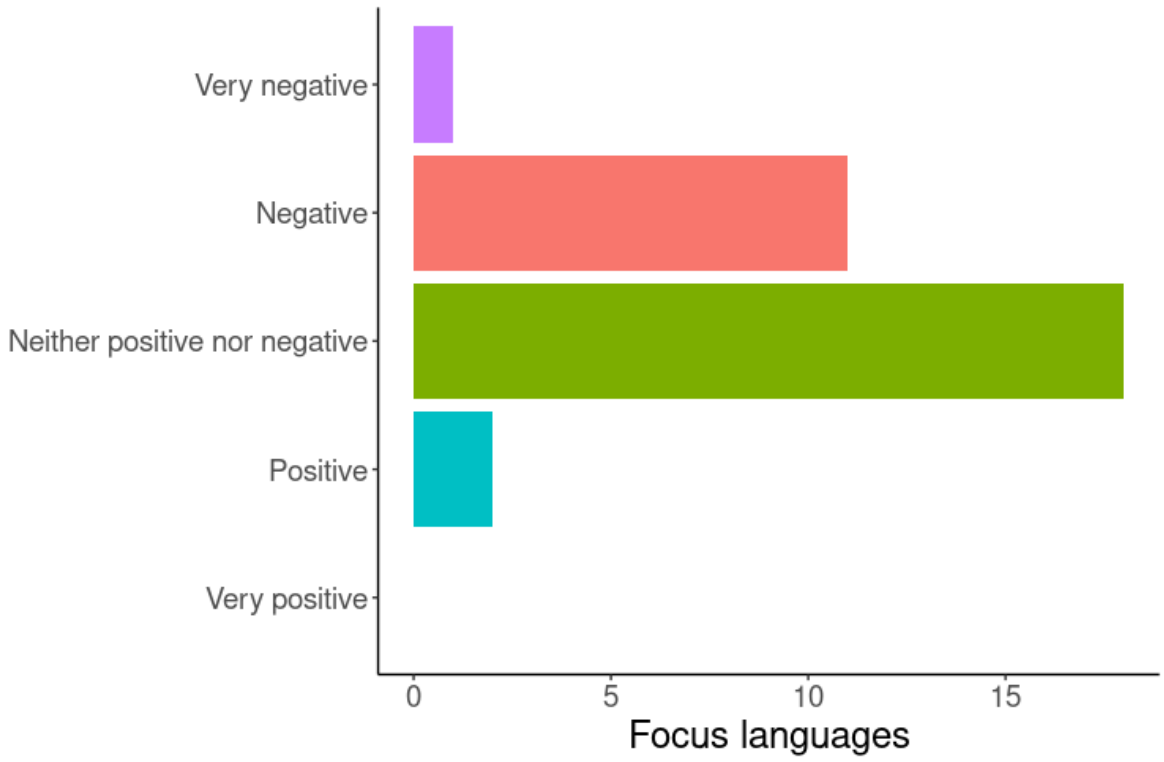


Figure 2: Distribution of responses about attitudes to linguistic transfer from the Neighbour Language to the Focus Language (data from Kashima et al. 2023).

Based on the specialists' responses, the Focus language speakers quite typically have either an indifferent or somewhat negative attitude to linguistic transfer from the Neighbor language. In other words, based on elicited specialists' assessments, it

appears that Focus language speakers would rarely consider lexical or grammatical borrowings from the Neighbor language as a good thing, although this response is not altogether absent in the data either.

As for the context-dependent factors of social contact, we asked questions about the social context of the contact across six social domains. The six social domains used are shown in (10) (see Kashima et al. 2025 for further details).

(10) List and definition of the GramAdapt social domains:

- Trade: concerning transaction of goods;
- Family and Kin: concerning relationships within the household;
- Local Community: concerning relationships in the private sphere beyond the household;
- Social Exchange: concerning relationships of non-transactional exchange, for example practices of ceremonial exchange;
- Knowledge: concerning relationships in the sphere of formalized learning (schooling and religion being the prototypes);
- Labor: concerning relationships in the sphere of production.

In most cases the same questions were repeated across all six social domains. This helped us to get an understanding of how the social aspects of contact between the Focus language speakers and Neighbor language speakers varied across social contexts.

We illustrate this point by taking another example from Kashima et al. (2023) that shows how the context-dependent aspect of social contact was built into a question asking about the occurrence of contact between the Focus and Neighbor group people. Instead of asking whether there was contact between the Focus and Neighbor people to begin with, we asked whether there was contact between the Focus and Neighbor people in each of the social domains. (11a), adapted to our example set Alorese-Adang in (11b), illustrates what this question looks like.

(11) Implementing the context-dependent aspects of social contact

a. Is there social contact between the Focus and Neighbor people in the domain of trade/labor/

b. Is there social contact between Alorese and Adang speakers in the domain of trade/labor/…

As is clear from this example, we often elicited the context-dependent and relational aspects of contact through one and the same question; context-dependency stems from anchoring the question in the individual social domains whereas relationality is related to the fact that contact is always framed from the perspective of the Focus-Neighbor interaction.

Figure 3 summarizes the occurrence of contact across the six social domains for 34 Focus-Neighbor pairs for which we have data on (Kashima et al. 2023). Based on the distribution of responses in Figure 3, there seems to be some evidence for a hierarchy of bilingual interaction across social domains. The hierarchy is summarized in (12).

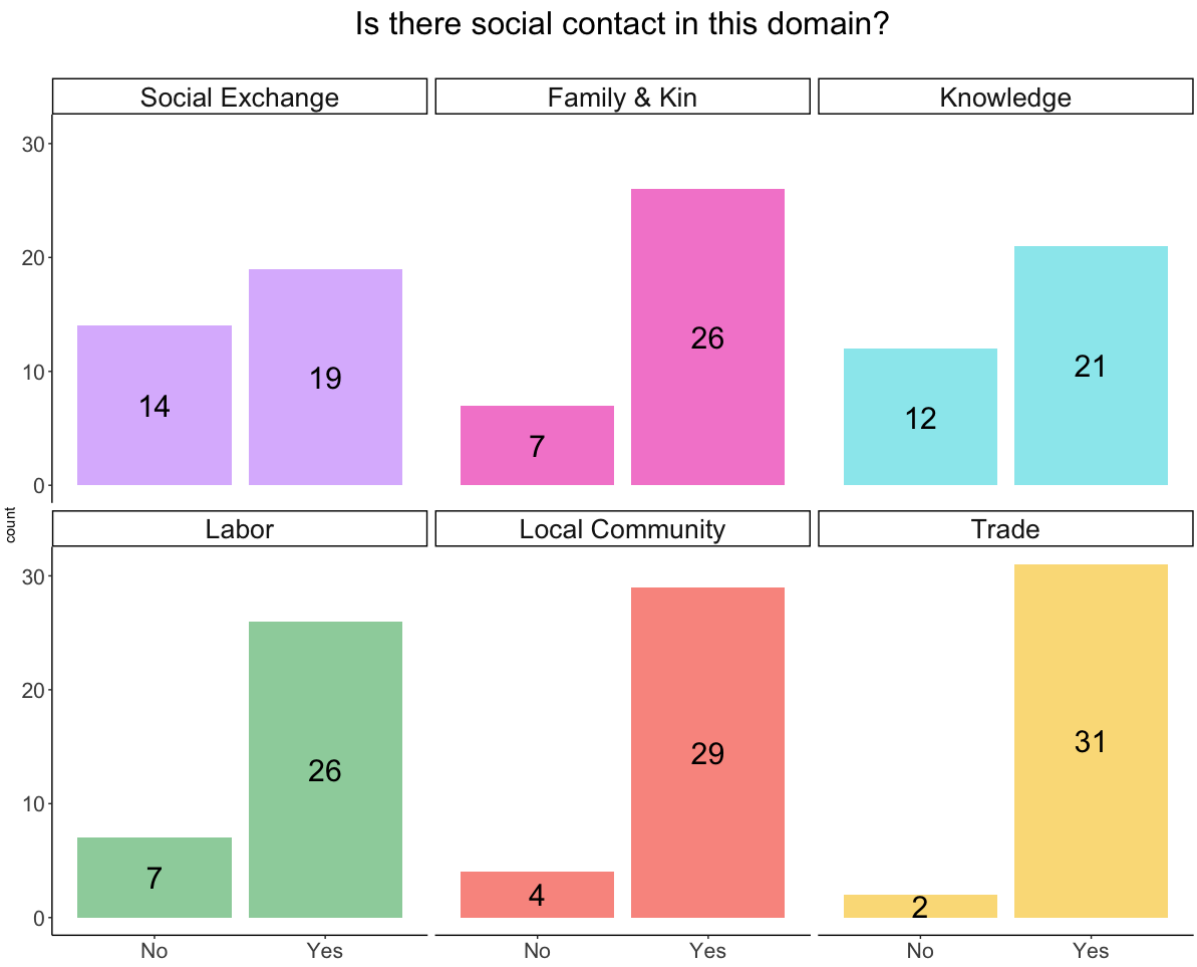


Figure 3: Distribution of responses on social contact between the Focus and Neighbour across social domains (data from Kashima et al. 2023).

(12) TRADE ≥ LOCAL COMMUNITY ≥ FAMILY & KIN ≥ LABOUR KNOWLEDGE ≥ SOCIAL EXCHANGE
97% 94% 82% 91% 82%

What this hierarchy states is that if there is contact in the domain of social exchange, there is likely contact in the other domains up in the hierarchy as well. However, if there is contact in the domain of trade, that does not yet tell whether there is contact in any other domains down the hierarchy. In our dataset, contact is thus most likely to occur in the domain of trade, which may also suggest that this is the context where contact relations between different language communities first develop. Conversely, language contact in the domains of knowledge and social exchange is the least typical and seems to first require contact in at least some other social domains higher in the hierarchy. The figures below the hierarchy reflect the percentage of the sampled sets in which the predictions of the hierarchy hold. For instance, only in six sets out of 34 are Focus and the Neighbor communities in contact in the domain of social exchange without being in contact in the knowledge domain. In other words, the prediction of the hierarchy for those two domains, i.e. contact in social exchange only if contact in knowledge exchange, holds in 27 out of 33 sets (≈ 82%). Overall, the predictions of the hierarchy hold to a very high degree (from 82% to 97%).

The sociolinguistic dataset stemming from our questionnaire responses contains data on roughly 200 variables on each set, totaling roughly 12,000 datapoints. The dataset makes it possible to answer various questions related to language contact and its social foundations, including questions about a particular contact scenario but also across contact scenarios. The list in (13) is a selection of possible general questions that can be potentially asked through this questionnaire.

(13) A sample of questions enabled by the GramAdapt Social Contact dataset

- Which sociolinguistic factors contribute the most to contact-induced change (cf. Thomason & Kaufmann 1988)?
- Does the rate of language change in contact situations depend on the social domain(s) in which contact takes place (cf. Greenhill et al. 2018)?
- Does the importance of language to group identity lead to greater divergence between languages (Braunmüller et al. 2014)?
- Does population size have any effect on contact-induced change and if yes, is it indirect so that it approximates the effect of some other social factors (Sinnemäki & Di Garbo 2018)?
- Is it possible to predict which sociolinguistic factors lead to simplifications or complexifications in contact situations (cf. Trudgill 2011)?

- Which social aspects of contact together affect contact-induced change (Trudgill 2011)?
- Does the extent of contact effects vary depending on whether children participate in contact dynamics (Trudgill 2011)?
- Are contact-induced changes in morphology driven more by changes in syntax than by aspects of the bilingual language ecology (Sinnemäki 2020)?

In the case of the Alorese-Adang contact pair which we exemplified in the previous section, sociolinguistic and historical data collected through the questionnaire (Moro & Sulistyono forthcoming) suggest that two main types of contact dynamics might have contributed to shape the patterns of contact-induced change attested in Alorese under the influence of Adang. Under the first scenario, possibly ongoing until the Dutch colonial period, the Adang dominated the Alorese. The Alorese-speaking population was bilingual in Adang and this situation of symmetrical bilingualism likely led to the patterns of grammatical restructuring discussed in 3.1, such as the grammaticalization of the plural word *hire* or the restructuring of adnominal possession patterns. After the start of the Dutch period, the Alorese prestige increased, which contributed to reshaping the extent of bilingualism from symmetric child-based bilingualism to Adang's adult-based bilingualism in Alorese. Language change processes stemming from this later type of contact settings mostly affect Alorese verbal morphology, which is undergoing considerable simplification under the influence of Adang L2 speakers (see Moro & Sulistyono forthcoming and references therein). This example clearly illustrates how the sociolinguistic data collected through the questionnaire may help elucidating the sociohistorical correlates of linguistic change and how these unfold through the history of a speech community.

4. Concluding remarks

In this paper, we tackled the broad question of how to define and account for "naturally occurring" data in linguistic typology by discussing the social anchoring of large-scale comparative research on language structures. We proposed that focusing on the relational and context-dependent properties of language is one way to investigate the social correlates of the distribution of linguistic diversity, thus ultimately boosting the social anchoring of typological generalizations.

We illustrated this point by presenting the research methodology of the GramAdapt project which has recently proposed a new way of approaching the dynamics of language change and linguistic diversification of the world's languages, focusing on language contact as a case in point. This novel approach intervenes on all key aspects of crosslinguistic research, from language sampling to data coding and statistical testing.

As shown in early sections of this paper, the GramAdapt typological and sociolinguistic datasets consist of observational data based on expert judgments, as is still typical of large-scale comparative research on linguistic diversity. Yet, the toolkit developed by this project strives to account for the relational and context-dependent aspects of language use at all stages of data collection and analysis. The sampling unit is eminently relational as it is constructed around documented contact scenarios between Focus and Neighbor languages, on the one hand, and proven genealogical relations between Focus and Benchmark languages, on the other. In addition, linguistic and sociolinguistic data are collected with the aim of capturing context-dependencies, both in terms of language-internal variation in the occurrence of individual linguistic features, and in terms of sociolinguistic variation with respect to observed dynamics of language use, language attitudes and ideologies across social domains.

Our approach to language variation and change is that of dynamic typology, whereby language universals are understood as universal pressure for type change. The time window for testing transition probabilities between linguistic types is the one provided by the comparative method. If we see a bias or a preference within that window, we can then extrapolate outside that window, other things being equal (Bickel 2013). But how justified is this extrapolation? In other words, have the factors that pressure languages to change remained the same even within this time window?

Languages are spoken in certain sociohistorical contexts, and many of those contexts have changed dramatically over the past few millennia (invention of agriculture and sedentary lifestyle, industrialization, urbanization, mass literacy, population movements, etc.). Research in (post-)colonial contexts squarely point to the coercive powers of colonial state infrastructure as responsible for societal upheavals (e.g. Givón 1971; vid. Yakpo 2020) which consequently act as pressures leading to language shift, death, and possibly the emergence of creoles and mixed languages. The linguistic consequences brought about by (forced) use of a state-sanctioned language, (forced) literacy and (forced) movement of people are well documented across the globe. We know less, however, about linguistic changes that

occur in non-colonial contact situations, and we are certainly still in the early days of developing a global-historical understanding of what types of pressures beget what kinds of linguistic changes – if there is indeed such a clear relationship. Our method partially allows us to tackle such future endeavors with the limits of its design principles.

Finally, the goal of the methodological shift that our methods propose and instantiate is not only that of getting new, groundbreaking answers to questions about the nature of human linguistic behavior. What our method also tries to achieve is an empirical test of how typological generalizations may change (or not) when integrating an explicit account of social ecology in our models. While genealogical and geographical bias control has been the only way to account for the social anchoring of language structures in typological models, our methods offer a much wider battery of hitherto largely untested factors, ranging from patterns of language transmission to attitudes and ideologies, which are anchored to social practices across individual domains of interaction.

Acknowledgements

Parts of this article were presented by Kaius Sinnemäki at the Second Workshop of the Nordic Signed Language Corpus Network (NSLCN) at Jyväskylä on 9-10 February 2023, by Francesca Di Garbo, Eri Kashima, Ricardo Napoleão de Souza, and Kaius Sinnemäki at the conference *Naturally Occurring Data in and beyond Linguistic Typology* at Bologna on 18-19 May 2023, and by Eri Kashima, Francesca Di Garbo, and Oona Raatikainen at the *Annual Meeting of the Societas Linguistica Europaea 55* on 8 August 2022. We are grateful to the audiences of these events for their comments, and to two anonymous reviewers for their constructive feedback on an earlier version of this manuscript. This research has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No 805371; PI Kaius Sinnemäki).

Abbreviations

1 = 1st person
3 = 3rd person
DEM = demonstrative
GEN = genitive
OBV = obviative
PL = plural
POSS = possessive
PROX = proximal/proximate
SG = singular

References

Adamou, Evangelia. 2021. *The adaptive bilingual mind: Insights from endangered languages*. Cambridge: Cambridge University Press.

Bell, Allan. 2013. *The guidebook to sociolinguistics*. Oxford: John Wiley & Sons.

Bentz, Christian & Bodo Winter. 2013. Languages with more second language learners tend to lose nominal case. *Language Dynamics and Change* 3(1). 1–27. https://doi.org/10.1163/22105832-13030105.

Berdicevskis, Aleksandrs. 2020. Foreigner-directed speech is simpler than native-directed: Evidence from social media. In *Proceedings of the Fourth Workshop on natural language processing and computational social science*, 163–172. Association for Computational Linguistics.

Berthele, Raphael. 2021. The extraordinary ordinary: Re-engineering multilingualism as a natural category. *Language Learning* 71(S1). 80–120. https://doi.org/10.1111/lang.12407.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Bickel, Balthasar. 2007. Typology in the 21st century: Major current developments. *Linguistic Typology* 11(1). 239–251. https://doi.org/10.1515/LINGTY.2007.018.

Bickel, Balthasar. 2013. Distributional biases in language families. In Balthasar Bickel, Lenore A. Grenoble, David A. Peterson & Alan Timberlake (eds.), *Language typology and historical contingency: In honor of Johanna Nichols*, 415–444. Amsterdam: John Benjamins.

Braunmüller, Kurt Steffen Höder & Karoline Kühl (eds). 2014. *Stability and divergence in language contact: Factors and mechanisms* (Studies in Language Variation 16). Amsterdam: John Benjamins.

Chen, Su-Chiao. 1997. Sociology of language. In Nancy H. Hornberger & David Corson (eds.), *Encyclopedia of language and education: Research methods in language and education* (Encyclopedia of Language and Education, vol. 8), 1–13. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-011-4535-0_1.

Cysouw, Michael. 2011. Understanding transition probabilities. *Linguistic Typology* 15(2), 415–431. https://doi.org/10.1515/lity.2011.028.

Cysouw, Michael & Jeff Good. 2013. Languoid, doculect and glossonym: Formalizing the notion ‘language’. *Language Documentation & Conservation* 7. 331–359.

Correia, Fabrice & Benjamin Schnieder (eds.). 2012. *Metaphysical grounding: Understanding the structure of reality*. Cambridge: Cambridge University Press.

Croft, William. 2001. *Explaining language change*. Harlow, UK: Pearson Education Limited.

Cronhamn, Sandra. 2018. *Quantifying loanwords*: *A study of borrowability in the Finnish lexicon*. MA thesis, Lund University, Lund.

Di Carlo, Pierpaolo & Jeff Good. 2023. Language contact or linguistic micro-engeeneering? Feature pool, social semiosis and intentional language change in the Cameroonian Grassfields. *Linguistic Typology at the Crossroads* 3(1). 72–125. Special issue on: Language contact and non-convergent change: cases from Africa (edited by Pierpaolo Di Carlo & Pius Akumbu). https://doi.org/10.6092/issn.2785-0943/17231.

Di Garbo, Francesca, Eri Kashima, Ricardo Napoleão de Souza & Kaius Sinnemäki. 2021. Concepts and methods for integrating language typology and sociolinguistics. In Silvia Ballaré & Guglielmo Inglese (eds.), *Tipologia e Sociolinguistica: Verso un approccio integrato allo studio della variazione: Atti del Workshop della Societá Linguistica Italiana 20 settembre 2020*), 143–176. Milano: Officinaventuno. https://doi.org/10.17469/O2105SLI000005.

Di Garbo, Francesca & Ricardo Napoleão de Souza. 2023. A sampling technique for worldwide comparisons of language contact scenarios. *Linguistic Typology* 27(3). 553–589. https://doi.org/10.1515/lingty-2022-0005.

Di Garbo, Francesca & Panagiotis Kapellis 2025. Contact effects in nominal number systems: A world-wide survey. *Studies in Language*. https://doi.org/10.1075/sl.24019.dig.

Dingemanse, Mark, Francisco Torreira & Nick. J. Enfield. 2013. Is “Huh?” a universal word? Conversational infrastructure and the convergent evolution of linguistic items. *PLOS ONE* 8(11). e78273. https://doi.org/10.1371/journal.pone.0078273.

Döhler, Christian. 2018. *A grammar of Komnzo*. Berlin: Language Science Press. https://doi.org/10.5281/zenodo.1477799.

Dunn, Michael & Simon J. Greenhill, Stephen C. Levinson & Russell D. Gray. 2011. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473. 79–82. https://doi.org/10.1038/nature09923.

Fishman, Joshua A. 1965. Who speaks what language to whom and when? *La Linguistique* 1(2). 67–88.

Fleming, Luke. 2011. Name taboos and rigid performativity. *Anthropological Quarterly* 84(1). 141–164.

Floyd, Simeon, Giovanni Rossi & Nick J. Enfield. 2020. A coding scheme for recruitment sequences in interaction. In Simeon Floyd, Giovanni Rossi & Nick Enfield (eds.), *Getting others to do things: A pragmatic typology of recruitments*, 25–50. Berlin: Language Science Press. https://doi.org/10.5281/zenodo.4018372.

Garde, Murray. 2008. *Kun-dangwok*: "clan lects" and *Ausbau* in western Arnhem Land. *International Journal of the Sociology of Language* 191. 141–169. https://doi.org/10.1515/IJSL.2008.027.

Givón, Talmy. 1971. Linguistic colonialism and de-colonialisation: The school system as a tool of oppression. *Ufahamu: A Journal of African Studies* 1(3). 33–49. https://doi.org/10.5070/F713016373.

Griffiths, Devin. 2017. The comparative method and the history of the modern humanities. *History of Humanities* 2(2). http://dx.doi.org/10.1086/693325.

Greenberg, Joseph. 1978. *Universals of human language*. Stanford: Stanford University Press.

Greenhill, Simon J., Xia Hua, Caela F. Welsh, Hilde Schneemann & Lindell Bromham. 2018. Population size and the rate of language evolution: A test across Indo-European, Austronesian, and Bantu languages. *Frontiers in Psychology* 9. 576. https://doi.org/10.3389/fpsyg.2018.00576.

Gumperz, John J. & Robert Wilson. 1971. Convergence and creolization: a case from the Indo-Aryan/Dravidian border in India. In Dell Hymes (ed.), *Pidgnization and creolization of languages*, 151–168. Cambridge: Cambridge University Press.

Guzmán Naranjo, Matías & Gerhard Jäger. 2023. Euclide, the crow, the wolf and the pedestrian: distance metrics for linguistic typology. *Open Research Europe* 3. 104. https://doi.org/10.12688/openreseurope.16141.1.

Guzmán Naranjo, Matías & Laura Becker. 2022. Statistical bias control in typology. *Linguistic Typology* 26(3). 605–670. https://doi.org/10.1515/lingty-2021-0002.

Haan, Johnson. 2001. *The grammar of Adang: A Papuan language spoken on the island of Alor, East Nusa Tenggara – Indonesia*. Doctoral dissertation, University of Sydney, Sidney. http://hdl.handle.net/2123/6413.

Haig, Geoffrey, Stefan Schnell (eds.). 2023. *Multi-CAST: Multilingual corpus of annotated spoken texts*. Version 2311. Bamberg: University of Bamberg. multicast.aspra.uni-bamberg.de/#veraa (Accessed 2025.12.05).

Haugen, Einar. 1972. *Ecology of language*. Stanford: Stanford University Press.

Hawkins, John A. 2014. *Cross-linguistic variation and efficiency*. Oxford: Oxford University Press.

Janssen, Dirk, Balthasar Bickel & Fernando Zúñiga. 2006. Randomization tests in language typology. *Linguistic Typology* 10(3). 419–440. https://doi.org/10.1515/LINGTY.2006.013.

Kashima, Eri, Francesca Di Garbo, Oona Raatikainen, Rosnátaly Avelino, Sasha Beck, Anna Berge, Ana Blanco, et al. 2023. *GramAdapt social contact dataset*. Available online at: https://doi.org/10.5281/zenodo.7508054.

Kashima, Eri, Francesca Di Garbo, Ruth Singer & Olesya Khanina & Ruth Singer. 2025. The design principles of a sociolinguistic-typological questionnaire for language contact research. *Language Dynamics and Change* 15(1). 1–103. https://doi.org/10.1163/22105832-bja10035.

King, Brian W. 2019. *Communities of practice in applied language research: A critical introduction*. London: Routledge.

Klamer, Marian. 2012. Papuan-Austronesian language contact: Alorese from an areal perspective. In Nicholas Evans & Marian Klamer (eds). *Melanesian languages on the edge of Asia: Challenges for the 21st century*, 72–108. Honolulu: University of Hawai'i Press.

Kluge, Angela. 2017. *A grammar of Papuan Malay*. Berlin: Language Science Press.

Kusters, Wouter. 2003. *Linguistic complexity: The influence of social change on verbal inflection*. Doctoral dissertation, University of Leiden, Leiden.

Ladd, D. Robert & Roberts, Seán G. & Dediu, Dan. 2015. Correlational studies in typological and historical linguistics. *Annual Review of Linguistics* 1. 221–241. https://doi.org/10.1146/annurev-linguist-030514-124819.

Levshina, Natalia. 2019. Token-based typology and word order entropy: A study based on universal dependencies. *Linguistic Typology* 23(3), 533–572. https://doi.org/10.1515/lingty-2019-0025.

Levshina, Natalia. 2022. *Communicative efficiency: Language structure and use*. Cambridge: Cambridge University Press.

Levshina, Natalia, Savithry Namboodiripad, Marc Allassonnière-Tang, Mathew Kramer, Luigi Talamo, Annemarie Verkerk, Sasha Wilmoth, et al. 2023. Why we need a gradient approach to word order. *Linguistics* 61(4). 825–883. https://doi.org/10.1515/ling-2021-0098.

Lupyan, Gary & Rick Dale. 2010. Language structure is partly determined by social structure. *PLOS One* 5(1). e8559. https://doi.org/10.1371/journal.pone.0008559.

Lüpke, Friederike. 2022. *Contact, concord, classification: Noun class systems and categorisation in a multilingual area*. (Paper presented at the seminar of the Helsinki Diversity Linguistics Group, Helsinki, 29 April 2022).

Lüpke, Friederike. 2024. Language, land and languaging in the Atlantic space. In Friederike Lüpke (ed.), *The Oxford guide to the Atlantic languages of West Africa*, 3–15. Oxford: Oxford University Press.

Mansfield, John, Henry Leslie-O'Neill & Haoyi Li. 2023. Dialect differences and linguistic divergence: A crosslinguistic survey of grammatical variation. *Language Dynamics and Change* 13(2). 232–276. https://doi.org/10.1163/22105832-bja10026.

Maslova, Elena. 2000. A dynamic approach to the verification of distributional universals. *Linguistic Typology* 4. 307–333.

Maslova, Elena & Tatiana Nikitina. 2007. Stochastic universals and dynamics of cross-linguistic distributions: The case of alignment types. (Stanford: Stanford University, Unpublished manuscript). http://anothersumma.net/Publications/Ergativity.pdf.

Merlan, Francesca. 1981. Land, language and social identity in Aboriginal Australia. *Mankind* 13(2). 133–148.

Milroy, James & Milroy, Lesley. 1985. Linguistic change, social network and speaker innovation. *Journal of Linguistics* 21(2). 339–384. https://doi.org/10.1017/S0022226700010306.

Moro, Francesca R. 2018. The plural word *hire* in Alorese: Contact-induced change from neighboring Alor-Pantar languages. *Oceanic Linguistics* 57(1). 177–198. https://doi.org/10.1353/ol.2018.0006.

Moro, Francesca R. 2021. Multilingualism in eastern Indonesia: linguistic evidence of a shift from symmetric to asymmetric multilingualism. *International Journal of Bilingualism* 25(4). 1102–1119. https://doi.org/10.1177/13670069211023134.

Moro, Francesca & Yunus Sulistyono. forthcoming. The Alorese and the Adang in eastern Indonesia. In Francesca Di Garbo & Eri Kashima & Kaius Sinnemäki (eds.), *Social foundations of language contact: A comparative survey*. Berlin: Language Science Press.

Muysken, Pieter. 2013. Language contact outcomes as the result of bilingual optimization strategies. *Bilingualism: Language and Cognition* 16(4). 709–730. https://doi.org/10.1017/S1366728912000727.

Nevalainen, Terttu & Helena Raumolin-Brunberg. 2017. *Historical sociolinguistics: Language change in Tudor and Stuart England*. 2nd edn. London: Routledge.

Nishiyama, Kunio & Herman Kelen. 2007. *A grammar of Lamaholot, eastern Indonesia: The morphology and syntax of the Lewoingu dialect.* München: Lincom.

Pakendorf, Brigitte Nina Dobrushina & Olesya Khanina. 2021. A typology of small-scale multilingualism. *International Journal of Bilingualism* 25(4). 835–859. https://doi.org/10.1177/13670069211023137.

Ross, Malcolm. 2013. Diagnosing contact processes from their outcomes: The importance of life stages. *Journal of Language Contact* 6(1). 5–47. https://doi.org/10.1163/19552629-006001002.

Rossi, Giovanni, Simeon Floyd & Nick J. Enfield. 2020. Recruitments and pragmatic typology. In Simeon Floyd, Giovanni Rossi & Nick J. Enfield (eds.), *Getting others to do things: A pragmatic typology of recruitments*, 1–23. Berlin: Language Science Press. https://doi.org/10.5281/zenodo.4018370.

Seifart, Frank Ludger Paschen & Matthew Stave (eds.). 2024. *Language documentation reference corpus (DoReCo)* 2.0. Berlin & Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2). Available online at: https://doi.org/10.34847/nkl.7cbfq779 (Accessed 2025.12.05).

Silverstein, Michael. 2003. Indexical order and the dialectics of sociolinguistic life. *Language & Communication* 23(3). 193–229. https://doi.org/10.1016/S0271-5309(03)00013-2.

Sinnemäki, Kaius. 2009. Complexity in core argument marking and population size. In Geoffrey Sampson, David Gil & Peter Trudgill (eds.), *Language complexity as an evolving variable*, 125–140. Oxford: Oxford University Press.

Sinnemäki, Kaius. 2010. Word order in zero-marking languages. *Studies in Language* 34(4). 869–912. https://doi.org/10.1075/sl.34.4.04sin.

Sinnemäki, Kaius. 2014. Cognitive processing, language typology, and variation. *WIREs Cognitive Science* 5(4). 477–487. https://doi.org/10.1002/wcs.1294.

Sinnemäki, Kaius & Francesca Di Garbo. 2018. Language structures may adapt to the sociolinguistic environment, but it matters what and how you count: A typological study of verbal and nominal complexity. *Frontiers in Psychology* 9. 1141. https://doi.org/10.3389/fpsyg.2018.01141.

Sinnemäki, Kaius. 2020. Linguistic system and sociolinguistic environment as competing factors in linguistic variation: A typological approach. *Journal of Historical Sociolinguistics* 6(2). 1–39. https://doi.org/10.1515/jhsl-2019-1010.

Sinnemäki, Kaius & Ahola, Noora. 2023. Testing inferences about language contact on morphosyntax: A typological case study on Alorese–Adang contact. *Transactions*

of the Philological Society 121(3). 513–545. https://doi.org/10.1111/1467-968X.12284.

Sinnemäki, Kaius, Francesca Di Garbo, Mark Ellison & Ricardo Napoleão de Souza. 2024. A typological approach to language change in contact situations. *Diachronica* 41(3). 379–413. https://doi.org/10.1075/dia.23029.sin.

Sinnemäki, Kaius. (in revision). On the "socio" in sociolinguistic typology: A review.

Sinnemäki, Kaius & Eri Kashima (forthcoming). Comparative historical sociolinguistics. In Terttu Nevalainen, Bridget Drinka, & Gijsbert J. Rutten (eds.), *Handbook of historical sociolinguistics*. Berlin: De Gruyter Mouton.

Suokhrie, Kelhouvinuo. 2016. Clans and clanlectal contact: Variation and change in Angami. *Asia-Pacific Language Variation* 2(2). 188–214. https://doi.org/10.1075/aplv.2.2.04suo.

Sulistyono, Yunus. 2022. *A history of Alorese (Austronesian) Combining linguistic and oral history*. Doctoral dissertation, University of Leiden, Leiden.

Thomason, Sarah G. & Terrence Kaufman. 1988. *Language contact, creolization, and genetic linguistics*. Berkeley: University of California Press.

Tomasello, Michael. 2010. *Origins of human communication*. Cambridge, MA: MIT Press.

Trudgill, Peter. 2011. *Sociolinguistic typology: Social determinants of linguistic complexity*. Oxford: Oxford University Press.

Wenger, Etienne. 1998. *Communities of practice: Learning, meaning, and identity*. Cambridge: Cambridge University Press.

Winter, Bodo, & Grice, Martine. 2021. Independence and generalizability in linguistics. *Linguistics 59*(5). 1251–1277. https://doi.org/10.1515/ling-2019-0049.

Witzlack-Makarevich, Alena, Johanna Nichols, Kristine A. Hildebrandt, Taras Zakharko & Balthasar Bickel. 2022. Managing AUTOTYP data: Design principles and implementation. In Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller & Lauren B. Collister (eds.), *The open handbook of linguistic data management (Open Handbooks in Linguistics)*, 631–642. Cambridge, MA: The MIT Press. https://doi.org/10.7551/mitpress/12200.003.0061.

Yakpo, Kofi. 2020. Social factors. In Evangelina Amadou & Yaron Matras (eds.), *The Routledge handbook of language contact*, 129–146. London: Routledge.

Zeman, Daniel et al. 2023. *Universal Dependencies 2.13*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of

Mathematics and Physics, Charles University. Available online at: http://hdl.handle.net/11234/1-5287.

CONTACT

francesca.di-garbo@univ-amu.fr
kaius.sinnemaki@helsinki.fi
eri.kashima@helsinki.fi or eri.kashima@amu.edu.au

Linguistic Typology at the Crossroads 5-2 (2025): 36-65 ISSN 2785-0943

For a *Discourse-Sensitive Typology*: Theoretical and methodological aspects

SIMONE MATTIOLA
UNIVERSITY OF PAVIA

Submitted: 09/04/2025 Revised version: 06/10/2025
Accepted: 06/11/2025 Published: 25/02/2026

Abstract

The aims of this paper are twofold. First, I aim to discuss some theoretical and practical issues concerning the methods employed by linguists when doing typological research. More specifically, I will present the kinds of data typologists usually base their work on, also discussing some related issues that deal with the nature of the data themselves and the possible misinterpretations that they may lead to. Then, I will discuss how typology tends to focus its attention over a limited array of phenomena (i.e., morphosyntactic ones), leaving aside phenomena pertaining to other levels of analysis (e.g., discourse). I thus will exemplify how integrating discourse phenomena might result in more comprehensive analyses and, consequently, in better-suited typological generalizations. Second, I will propose a multi-level method (which I call *discourse-sensitive typology*) based on the *converging evidence* perspective, discussing how it might help us to overcome the abovementioned issues and ultimately make typology truly usage-based.

Keywords: typology; discourse; methods; usage-based.

1. Introduction

This paper aims at discussing some theoretical and practical issues concerning the methods employed by linguists when doing typological research, intended here as cross-linguistic comparison of linguistic phenomena both at the large-scale worldwide

and micro-typological level (i.e., on a specific geographic area or genealogical groupings of languages), proposing a different view and method (that I call *discourse-sensitive typology*). Comparing languages is not an easy task under several respects, and typological literature has, in some way, dealt with some of the issues that doing research in this field has raised, often finding very good solutions or at least creating vivid debates within the scientific community. Since its (recent) foundation (in the 1960s by Joseph Greenberg, at least in its modern sense), typology has incredibly developed its theoretical bases in all the different nuances, reaching a prominent position within linguistics. In fact, on the one hand, we assisted at remarkable theoretical advancements on how to define phenomena cross-linguistically (e.g., comparative concepts and the debate around them, see Dryer 1997; Croft 2001, 2003, 2022; Haspelmath 2007, 2010; Cristofaro 2009; among many others) or on how to explain typological generalizations (see e.g. Schmidtke-Bode et al. 2018), but also in terms of methodological tools (e.g., quantitative typological approaches, see for example the notions of *distributional typology* and *multivariate analysis* in typology as in Bickel 2015 and Bickel 2010, 2011 respectively, but also the methodological apparatus in Guzmán Naranjo & Becker 2022 and in Becker & Guzmán Naranjo 2025, just to mention a few of them); on the other hand, we still have some of these aspects that have basically been neglected or at least for which scarce advancements have been made. Among these, the reflections on the data and the phenomena on which typology bases its analyses remained underexplored. Even though I do not believe in distinguishing between qualitative and quantitative research, at least in typology (since every research has some qualitative and quantitative nuances at the same time), we can say that advances have been made for what concerns the methods in "quantitative" typology, as just mentioned, but "qualitative" typology remains stuck in its position for what concerns the methodological debate.

After introducing some preliminaries in section 2, section 3 aims at discussing the types of data and some related issues for typological research. First, the different possible sources that typologists usually work on are listed and presented (section 3.1) showing how these can be problematic for the typological analysis (section 3.1.1). Second, I briefly discuss the phenomena on which typologists focus their attention (section 3.2) exemplifying how typology could benefit from also looking beyond at less canonical phenomena (section 3.2.1). In section 4, I propose a multi-level method to overcome the issues discussed in previous sections and that would allow typology to become a discourse-sensitive field. Section 5 draws some conclusions.

2. Preliminaries for a *discourse-sensitive typology*: the *why* and the *how*

As already pointed out, the goal of this paper is to propose a discourse-sensitive typology discussing the role that discourse should acquire in typology both in terms of discourse-informed analyses of grammatical phenomena and of linguistic analyses of discourse phenomena.

Linguistic typology is usually defined as "the study of linguistic systems and recurring patterns of linguistic systems" allowing typologists to identify universals which "are typological generalizations based on these recurring patterns" (Velupillai 2012: 15). Croft (2003) identifies at least three different senses that the term typology can have in linguistics: (i) typological classification, (ii) typological generalization, and (iii) typological approach which he defines as follows:

1. ***Typological classification***: "a classification of structural types across languages. […] [A] language is taken to belong to a single type, and a typology of languages is a definition of the types and an enumeration or classification of languages into types" (Croft 2003: 1).
2. ***Typological generalization***: "the study of patterns that occur systematically across languages. […] The patterns found in typological generalization are language **universals**" (Croft 2003: 1, bold in the original).
3. ***Typological approach***: "typology represents an approach or theoretical framework to the study of language that contrasts with prior approaches, such as American structuralism and generative grammar" (Croft 2003: 2).

These three senses represent the three phases of the scientific method: collect and classify the data (*typological classification*); analyze the data, identify possible recurrent patterns and propose generalizations (*typological generalizations*); and account for some patterns which are cross-linguistically recurrent and why generalizations are found (*typological approach*). Croft (2003: 2, bold in the original) himself recognizes typology as "an **empirical scientific** approach to the study of language". As a natural consequence of its strong empirical foundation and its functionalist approach (see Croft 2003: 2), in the last decades, the usage-based approach has acquired more and more relevance in typological(-functionalist) literature. According to this approach, grammar is shaped by (and in some way it adapts to) usage in discourse (cf. Givón 1979a, Du Bois 1985, Bybee & Beckner 2010,

Diessel 2019, among many others). In other words, as Bybee (2006: 730) puts it, "[u]sage feeds into the creation of grammar just as much as grammar determines the shape of usage". This usage-based perspective allowed linguists from different fields to re-consider some grammatical phenomena, like, for example, syntactic ones (see the notion of *spoken syntax* as proposed by Hopper 1987, 1988). However, this perspective has not been fully adopted by all the approaches to the study of language and grammar, including some that share this functionalist view.[1] So far, these approaches have not (at all or not much) paid the required attention to spoken language and to phenomena that mainly pertain with the discourse level, and this is because they have mainly focused on "grammar", leaving aside "discourse". Linguistic typology is one of these disciplines. This is in some way surprising, and I think typologists should address this issue to make typology truly usage-based.

3. What data and what phenomena for typology

The situation described in section 2 seems to originate from two main reasons (at least in "qualitative" typology): (i) the lack of a full awareness of the kind of data typologists usually analyze, and (ii) the array of phenomena that typologists usually investigate. For what concerns the former, typology seems not to have always been fully aware of the nature of the data on which typological investigations are based; while, for the latter, typology, as already noted, tends to give particular attention to some levels of analysis (morphology and syntax) at the expense of others (e.g., phonology and discourse). Needless to say, these two issues are strictly connected with each other since the kind of data typologists analyze in some ways "mirrors" the phenomena investigated, and vice versa (see sections 3.1.1 and 3.2 below). The following sections will focus on these two issues, showing how they can be problematic and why they should be addressed.

3.1. Typological data and their issues

Traditionally, typologists retrieve their data from a range of different sources. Among the most commonly adopted, we can list the following: (i) questionnaires specifically

[1] Actually, this view was originally adopted in some typologically-oriented works by exponents of the so-called "West Coast functionalism", like Wallace Chafe, Talmy Givón, Sandra Thompson, and Marianne Mithun (see, e.g., Chafe 1976, 1987, 1994; Givón 1979a, 1979b, 1983, 1984; Mithun 1996, 2015; Thompson 1988), but it did not fully make it into large cross-linguistic investigations.

designed for the purposes of the investigation; (ii) parallel texts; (iii) dedicated scientific papers; and (iv) grammatical descriptions. In what follows, the merits and the imperfections for each of these kinds of data are briefly presented.[2]

Typological questionnaires

The first kind of data are questionnaires. Two sub-types can be identified: questionnaires to be submitted to native speakers and questionnaires to be submitted to linguists. The former is usually composed of a series of sentences to be translated by speakers from a meta-language (e.g., English) into the target language. An example is provided by Dahl's (1985) questionnaire for eliciting data for TAM categories composed of 156 sentences to be translated, as those reported in (1):

(1) Sentences 61-63 of Dahl's (1985) TAM questionnaire (Dahl 1985: 200-201):
   61. [It is cold in the room. The window is closed. Q:] You OPEN the window (and closed it again)?
   62. [Answer to (61):] (Yes,) I OPEN the window
   63. [Answer to (61):] (No,) I not OPEN the window

The sentences reported in this kind of questionnaire usually provide some context (e.g., in (1) the context is between squared brackets), and the uppercase form is the one of interest for the research. This form usually appears in its citation form in order to avoid biased answers from the informant towards the meta-language. So, for example, sentences 61-63 would be the following if our target language is Italian (ita; Indo-European, Italic):[3]

(2) Italian translation of Dahl's (1985) sentences 61-63 (own knowledge):
   61. [It is cold in the room. The window is closed. Q:] ***Apri** la finestra (e la richiudi)?*
   62. [Answer to (61):] *(Sì,)* ***apro** la finestra*
   63. [Answer to (61):] *(No,) non* ***apro** la finestra*

[2] The list is not intended as necessarily exhaustive, but it reports the most common types of sources and briefly discusses them.

[3] Language classification follows the one proposed in Glottolog 5.2 (Hammarström et al. 2025).

Thus, in this way, we would have data for the Italian Present Tense for the verb *aprire* 'open' and similarly for all the languages for which we have informants to answer the questionnaire.

The second type of questionnaire consists of a series of direct questions about the linguistic properties of the target language that the investigator submits to another linguist. For example, see (3):

(3) Questions 1-2 of Corbett's questionnaire on grammatical number (emphasis in the original):[4]

1. Which grammatical numbers are distinguished (singular-plural, singular-dual-plural, etc)?
2. How is number expressed?
   2.1 lexically: are there separate words meaning, say 'plural'? (It would be surprising to find such cases in Europe.)
   2.2 morphologically
      2.2.1 which means are used?
         inflectional: prefixing, suffixing, infixing, ambifixing
         other - suppletion, reduplication
      2.2.2 which lexical categories carry the morphological markers - nouns, verbs, adjectives, pronouns, others?
      2.2.3 within the lexical categories, are all items involved? (i.e. if adjectives mark number, do all or only some adjectives mark number?) It is common for nouns to be defective (singularia tantum, pluralia tantum etc); if so which are involved? Sometimes there are types of noun (e. g. abstracts, mass nouns) which can be predicted to be defective.
   2.3 syntactically
      2.3.1 is there a matching of number marking between different elements (especially in the NP) which could be characterized as agreement?
      2.3.2 if so, are there instances where such matching is violated (e.g. English The committee have decided, Norwegian Pannekaker er godt 'Pancakes is good')?
      2.3.3 are there particular syntactic complications where numerals are involved?

In this way, the researcher would have at her own disposal the required information on the phenomenon she is investigating from linguists who are expert of some languages.

Both these types of questionnaires have some merits but also some shortcomings. First, the questionnaires allow the gathering of data directly on the target phenomenon with specific information on the structures. On the one hand, the first

[4] This questionnaire can be found at the following website: https://www.eva.mpg.de/lingua/tools-at-lingboard/questionnaire/gender-and-number_description.php.

type (questionnaire for speakers) also allows the researcher to see how the specific phenomenon works in the language in controlled situations and specific pre-established contexts. On the other hand, the second type (questionnaire for linguists) allows to have a thorough and informed picture. Coming to the critical issues, the questionnaires can be filled out only by a limited number of people since it is difficult to submit it to large number of speakers/experts in a reasonable time (assuming that she is able to find enough of them) and, thus, the language sample will be relatively small (we can assume about 100 languages). In addition, the first type of questionnaire has a higher possibility to be submitted to informants of languages with a large number of speakers biasing the research towards WEIRD and LOL languages (WEIRD = Western, Educated, Industrialized, Rich, and Democratic, see Henrich et al. 2010 and Majid & Levinson 2010; LOL = Literate, Official, and Lot of users, see Dahl 2015). Then, this kind of data might suffer from influences by the meta-language due to the translation process: no matter how hard we try to avoid influences of this type, it is not possible to eliminate them. Also, in the first type of questionnaire, the context is *de facto* established by the researcher thus eventually biasing the research towards parameters of analysis adopted by the researcher herself. The second type of questionnaire does not provide actual data but rather analyses of data heavily dependent on the linguist's (possibly a non-typologist) own conceptualization of the phenomenon and thus without the possibility to check it through data analysis of any sort.

Parallel texts

The second kind of source typologists usually adopt are parallel texts. Cysouw & Wälchli (2007: 95) provide the following definition: "[p]arallel texts are texts in different languages that can be considered translational equivalent". They also introduce the notion of *massively parallel text* (MPT) "for such texts of which many different translations are available" (Cysouw & Wälchli 2007: 95). In other words, (M)PTs are different translations of the same text in different languages. Several cases of texts are largely translated, just to cite some of them: the Bible, *The Little Prince*, the Harry Potter saga, subtitles of movies, etc. Of course, this kind of data can be easily used to fulfill typological investigations, and they have both very good merits and some relevant shortcomings. First, MPTs allow the typologist to observe exactly the same occurrences of the phenomenon under investigation in the exact same co-text/context for different languages, and this represents a very strong quality. Second,

MPTs are usually digitalized and tagged texts that can be easily used for typological-oriented queries (and beyond). Third, we have at our disposal MPTs for several different languages according to the nature of the original text: from some dozen of languages (e.g., movie subtitles) to a few hundred (*Universal Declaration of Human Rights*, *The little prince* or Andersen's fairy tales), or even thousand (the Bible or part of it) (for a more precise count see Cysouw & Wälchli 2007). However, at the same time, MPTs are heavily biased by the translation since (often) the objective of the translator is not to provide a linguistically equivalent in two (or more) different languages, but rather to provide the more accurate semantic/sense equivalent of the intentions of the authors of the text. This issue can ultimately question if the "same" occurrences in different languages are actually the same, or they are rather "similar" occurrences with almost the same general meaning. This also opens another theoretical problem: are two structures and/or contextual/co-textual environments of two different languages equivalent at all? The answer is not easy to give and, at least for those who advocate for the non-existence of pre-established categories and constructions for typological research (see the "comparative concepts" debate), it usually seems to be "no".[5] However, this is a problem that holds for any kind of source, even though it might have a different impact on the analysis. Another relevant problem is that often parallel texts provide only the original text and its translation without any interlinear gloss making it difficult to pursue more fine-grained linguistic analyses (in particular, morphological and syntactic ones). Finally, MPTs are usually texts written in highly codified language, that is, they are mainly literary texts or, in the best-case scenario, texts that try to mirror spontaneous speech (e.g., subtitles).

Dedicated scientific papers

The third kind of data on which typologists can base their analyses are scientific papers. With this, I refer to papers written by linguists and appeared on scientific (international or not) journals which aim to describe a specific phenomenon in a single language or a group of languages (often from the same language family or geographic area). The main merit of this kind of data is that scientific papers tend to

[5] This position is of course debatable, but since this is not the objective of the present paper to discuss it here, I just refer to the main bibliographic references (already mentioned above in section 1) on it and the references cited therein to which I refer for further information (Dryer 1997; Croft 2001, 2022; Haspelmath 2007, 2010; Cristofaro 2009; among many others). For my purposes, it is only important to note the existence of different views on this that ultimately represent a possible issue.

be very specific about the object of analysis and to report (almost) all the information needed to account for a specific phenomenon in that/those language/s. Sometimes, the information that scientific papers provide is even too specific for typology, whose final aim is to simplify linguistic complexity to identify recurrent patterns and provide them with an explanation. Scientific papers also have a relevant issue for typological research: there will never be enough papers dealing with the object phenomenon for a sufficient number of languages. In other words, the number of languages for which we can find information in dedicated papers is not high enough to be accounted for as a(n actual) language sample. This kind of source can eventually be considered complementary support to the adoption of other sources.

Grammatical descriptions (or descriptive grammars)

The last type of data that typologists have at their disposal are grammatical descriptions (henceforth simply grammars). Grammars are probably the most used and the most traditional source for typological investigations. They consist of thorough descriptions of the grammatical structures of a language made by field linguists who collect texts (of different genres)[6] by speakers of the object language and provide an analysis of the structures they are able to identify in such a sample of texts.

As with all the other kinds of data, grammars have merits and shortcomings. The first merit is undoubtedly the fact that we have at our disposal grammatical descriptions of some kind for a large number of languages. According to Glottolog 5.2 (Hammarström et al. 2025), we have at least a full grammar or a grammar sketch for about 4.900 languages of the world, out of which at least ca. 2.900 languages have a full grammar. If we take for granted the overall number of about 7.800 languages of the world, we are talking about 63% of the total amount of languages.[7] Thus, this source would allow us to have incredibly extensive language samples (i.e., thousands of languages), a coverage that other kinds of data can hardly achieve. Another merit of grammars is that they provide a(n) (presumptive) overall description of the grammatical structures of a language, thus allowing to check for possible correlations among different constructions. Finally, one last value is that, quite often (at least in

[6] With the terms *text* and *genre*, I intend (here and elsewhere) them in their widest senses, that is, any possible type of linguistic text: written or oral, formal or informal, narrative or conversational, and so on.

[7] Of course, the figures are approximated for convenience, for the current figures see https://glottolog.org/langdoc/status.

grammars published in the last couple of decades), they also have a sample of (glossed) texts (at the end of the grammar) that allow us to look at the phenomenon/a we are investigating directly in texts. However, it is worth mentioning that, unfortunately, these sample texts often tend to be quite small.

Grammars also have some shortcomings. We can mention at least two of them that are relevant to the present discussion. First, the information we find within grammars are (at least) second-hand data, that is, they are not raw data to be analyzed by typologists, but they rather are data already analyzed by the grammarian (this is true also for other kinds of data, such as the one from questionnaires for linguists or dedicated scientific papers). This means that typologists must trust grammarians and their descriptions without debating much on what they find since they are not language experts. Of course, the grammarian might be contacted for further information, but this is not always possible or not for all the doubts that a typologist might have. Then, a follow-up of this shortcoming is that field linguists cannot have in mind all the questions that typologists would like to answer. This means that we cannot necessarily presume to find what we are looking for in a grammar or enough information on it and this does not necessarily mean that that specific language lacks the phenomenon we are investigating, but it simply means that the grammarian did not write anything about it because of a series of possible reasons such as the phenomenon is not relevant for the language or is categorized differently in the linguistic tradition of the language or any other possible reasons. All these reasons are valid and do not underestimate the immense work that fieldworkers do: typology would not exist without fieldwork. In the end, grammars cannot be considered by definition fully exhaustive for typological purposes, and this is because of the distinction between descriptive categories of individual languages and comparative concepts, two notions that are "in a many-to-many relationship" (Haspelmath 2010: 665) with each other.

The "perfect" source for typology

As already noted in the previous sections, none of the kinds of data presented above can be considered completely satisfactory. All of them show some shortcomings: some are more relevant or difficult to be faced, others are less. In any case, the fact that none of the sources is fully appropriate for the purposes of a thorough typological investigation clearly emerges from what has been shown above. This does not mean that typological research cannot be carried out, but rather that the data typologists analyze are not necessarily exhaustive. The most important consequence is that we

(typologists) must be aware of this and bear it in mind, in particular when proposing generalizations. Otherwise, typological data might bias our generalizations and, more in general, our research.

Picking one of these sources instead of the others is just a matter of personal choice and preference or of better adaptation of the source to the aims of our research. It is beyond the purposes of the current paper to discuss which is the best source of data, but I think that a personal positioning is in order in a paper like the current. In my opinion, descriptive grammars remain the best choice for several kinds of typological investigations and this is because they tend to be free from translational biases (much more than other sources) and pay the right tribute to language experts and field linguists who, by describing the languages, disclose their access to typologists thus allowing us to do cross-linguistic investigations.

The "perfect"[8] source would be to have at our disposal a significant corpus of (glossed) texts for a very large number of languages of the world, alongside a grammatical description. In this way, we would be able to investigate our object of analysis directly in the texts of the given language (preferably different textual genres) also having its context of use and co-text. In addition, the grammatical description would serve as a complementary tool to decipher and better understand the structures of a language in which we are not experts. Unfortunately, to date, this source is not available. Yet, in the last decade or so, a few projects have been carried out aiming to fill this gap. I am referring to projects like: CorpAfroAs (see Mettouchi et al. 2010, Mettouchi et al. 2015) which is specifically dedicated to Afroasiatic (AA) languages and contains texts and grammatical information for 16 AA languages; DoReCo (Seifart et al. 2024) which is a collection of spoken language corpora for 53 typologically different languages; and MultiCAST (Haig & Schnell 2023) which collects comparable corpora of 20 languages. Even though these projects are extremely important and praiseworthy, the final goal is still far, and probably a lot of time will pass before having this kind of source for a large number of languages.

3.1.1 How typological data may affect typological analyses

Let us now turn to how the kind of data adopted might impact the typological analysis. In the great majority of the cases, typologists use data from grammatical descriptions (as already noted above): grammars, being the best available source or

[8] Since talking about "perfectness" in science is always rather utopistic, I decided to use this term only between double quotation marks.

not, are still the most widely adopted source for typological research. In the best-case scenario, in descriptive grammars, we find data that are out of their narrative and/or pragmatic context and generally used by the grammarian to explain and exemplify some kind of phenomena of the object language in the specific paragraph/chapter of the grammar itself.

The lack of context (and co-text) can give rise to a possible 'partial' reading of the data that can ultimately lead to a possible misinterpretation of the data themselves. To interpret at best examples of phenomena found in grammars, the researcher should infer what in cognitive linguistics is called *construal*, that is, how "an experience is framed" in the sense of "how the speaker conceptualizes the experience to be communicated, for the understanding of the hearer" (Croft & Cruse 2004: 19). The narrative context/co-text[9] is indeed part of the construal since it would determine the conceptualization and the understanding of the linguistic structure (in its widest sense). From this, it follows that having an example of a specific phenomenon without its narrative context/co-text is like having incomplete information that ultimately can affect the interpretation. This, in turn, can lead to a possible biased typological generalization (if not even wrong): if typologists analyze a single (or a few) example(s) for the phenomenon they investigate in a specific language and this example might be wrongly interpreted given the lack of narrative context/co-text, then it follows that the overall understanding of the phenomenon might eventually be biased. Needless to say, this is a key issue for typology.

In the remainder of this section, I will illustrate through a couple of examples, based on the analysis on the pluractional marker *-pödï* of Akawaio (ake; Cariban, Venezuelan Cariban) in Mattiola & Gildea (2023), how this issue might impact typological analyses. With pluractional marker (PM), I intend any morphological strategy encoding a plurality of the event expressed by the verb (cf. Mattiola 2019: 164). Cross-linguistically, PMs can express a wide range of different functions, going from iterative to habitual and from durative to reciprocals (see Mattiola 2019: 21-42). As for any typological work, identifying the correct function for a formal strategy is fundamental. As noted above, to interpret the function of a strategy, we need to understand the construal of the construction and, specifically for PMs, the *event construal*: how the event is conceptualized and then expressed from a linguistic point

[9] With *narrative context*, I refer to the general context in and through which information (narration, in its widest sense) flows to be communicated. With *narrative co-text*, I intend the particular linguistic context in which a specific construction is inserted.

of view. This means we need to analyze the semantic-functional value that the PM adds to the lexical value of the verb and how it impacts the overall semantics of the example.

The PM *-pödï* of Akawaio can express several pluractional functions that we found in other languages of the world (Mattiola & Gildea 2023: 463-470). Let's now try to understand the functional reading of this marker in a couple of examples. Consider (4).

(4) Akawaio (Cariban, Venezuelan Cariban; Caesar-Fox 2003: 336)
*yöi naga'pï po y-enggurumï-**bödï**-Ø-ng*
tree stump on 3-wait-**PLAC**-PRES-STYLE
'He **would just rest** there on top of a piece of tree stump'

In (4), we may consider *-pödï* as expressing a continuative function ("a single situation that is prolonged during a period of time" Mattiola 2019: 34): the subject ("he") seems to be in the state of resting on the tree stump for a long time. Another possible reading might be to understand *-pödï* to express a wish, a sort of optative marker. However, this interpretation contrasts with the cross-linguistic understanding of PMs and thus appears to be less likely. But if we consider the narrative context the picture changes. The sentence in (4) is taken from a narrative story in which the main characters are a duck and an owl. Every day, both of them leave their own houses, but while the duck goes outside to go to work, the owl goes outside to rest on a tree stump waiting for the end of the day. Thus, a different interpretation seems to consider the state of resting as repeated over a long past time frame ("every day the owl rests on a tree stump"), that is, a habitual function ("a situation that is repeated customarily, i.e. that is typical of a period of time" Mattiola 2019: 31-32).

A similar situation is found when interpreting the sentence in (5).

(5) Akawaio (Cariban, Venezuelan Cariban; Caesar-Fox 2003: 522)
*ö'rö pe y-eji-Ø y-aburö-**bödï**-Ø*
what ATTR 3-be-PRES 3-praise-**PLAC**-PRES
'Why **is** she **being praised**?'

In this case, the PM might be interpreted as giving again a continuative reading or, eventually, an iterative reading to the situation ("the situation occurs multiple times,

but the repetitions are limited to a single and the same occasion" Mattiola 2019: 23). The sentence is a question in which the speaker asks the reasons why a girl/woman is being praised and we might interpret it as the action occurs several times (iterative) and/or prolonged through time (continuative). However, the example is found in a personal narrative in which the informant is reporting a party for the retirement of a nurse. Again, given the full contextual environment, the first-glance interpretation might be erroneous. Also in this case, it is more likely a different construal of the situation. In fact, a more well-suited interpretation seems to be participant plurality ("the plurality of situations will be distributed over different participants" Mattiola 2019: 26) or, but less likely, a frequentative reading ("the repetitions of a specific situation are performed over multiple and different occasions" Mattiola 2019: 24): several single events of praising in which several people congratulate the nurse for the retirement or a single participant (but also several) that congratulate with the nurse on different occasions during the party (frequentative). In any case, the first interpretation turned out to be wrong.

From these two very short examples, it clearly comes out the problem to which I was referring at the beginning of the current section: how data typologists collect and analyze might lead to possible misinterpretation if we do not take into consideration their narrative context. Again, this is a highly problematic issue because an incorrect (or wrong) interpretation of the data can result in a biased analysis and thus lead to incorrect (or wrong) generalizations.

In section 4, I will be back on how to try to avoid this issue with the data we have at our disposal.

3.2. *Investigate 'em all*: discourse phenomena in typological perspective

The second issue that deals with typology as a truly usage-based field concerns the range of phenomena typically examined by typologists. Often, typological investigations focus their attention on mainly "structural" and "grammatical" phenomena, that is, on morphology and syntax with phonetics/phonology and lexicon being much less represented in typological studies. Just as an example, suffice it to say that in the online version of the pioneering WALS project (Dryer & Haspelmath 2013), one of the most important sources and enterprise in the history of typology, out of the 144 chapters (to date), we can count 54 and 57 chapters respectively for morphosyntax (10 morphology, 28 nominal categories, 16 verbal categories) and

syntax (7 nominal syntax, 19 word order, 24 simple clauses, 7 complex sentences), 19 for phonology and 10 for lexicon (plus 2 for sign languages and 2 classified as "others", i.e., clicks and writing systems). The picture is thus quite straightforward: typology heavily prefers morphology and syntax over the other levels of analysis. On the one hand, we can, of course, account for this imbalance making reference to the historical relevance that morphology and syntax have for typology since its foundation. On the other, however, the imbalance is impressive and cannot be easily dismissed by referring to historical reasons and tradition. The picture becomes even harsher if we consider what is generally called "discourse phenomena", which, following Barotto & Mattiola (2023a), are defined as:

> linguistic elements and constructions that help to manage the organization, flow and outcome of communication (cf. Schiffrin 1987; Du Bois 2003). They have to do with what can be called information packaging, that is, the ways speakers organize their discourse and turns, link and give cohesion to utterances while clarifying the relationship occurring between them or foreshadow the status of information in what will come after. Moreover, discourse phenomena can also facilitate the intersubjectivity between speakers, as in the case of politeness or hedging. (Barotto & Mattiola 2023a: 1)

In fact, except for a few notable and recent cases (e.g., several papers on different phenomena in Barotto & Mattiola 2023b; Dingemanse 2012 and Lahaussois et al. 2024 on ideophones; Lahaussois & Treis 2019 and Ponsonnet et al. 2023 on interjections; Pakendorf & Rose 2025 on fillers), typology has almost completely ignored discourse phenomena. Studies focusing on discourse phenomena are pretty rare both at the cross-linguistic level and at the language-specific level for non-WEIRD and non-LOL languages. This is something that is not in line with the view that I gave in section 2, for which typology understands itself as a usage-based field (grammar is shaped by discourse use). In my opinion, we (typologists) must realize that giving the right importance to all the levels of analysis and in particular to discourse phenomena (which are by no means less relevant than morpho-syntactic ones) and their cross-linguistic variability can provide important information allowing us to better understand how languages shape their own grammar and why they are organized in such a way. Of course, this is not simple at all, and I will be back on how to do this in section 4. In the next section, I am going to show how typology (and discourse studies) can benefit from studying discourse phenomena in cross-linguistic

perspective and/or in non-WEIRD and non-LOL languages. To do so, I will briefly analyze a discourse construction found in Akawaio (Cariban, Venezuelan Cariban).

3.2.1 A discourse marker in a non-WEIRD/non-LOL language: ti'tuik prarö *in Akawaio*

Akawaio is a variety of Kapóng,[10] a Cariban language belonging to the Venezuelan branch of the family (Pemón sub-branch), spoken by the Akawaio tribe (circa 10.000 people) in Guyana, South America. The corpus I consulted (found as an appendix to Caesar-Fox 2003) is composed of about 10.800 words and was collected, analyzed, and glossed by Desrey Caesar-Fox (with Spike Gildea). The corpus is composed of twenty-seven texts belonging to different genres: traditional stories (12), personal narratives (5), Tareng healing chants (6), and traditional praising rhymes for children (4). In this corpus, I came across a construction, *ti'tuik prarö*, whose formal and functional properties revealed to be peculiar. For this reason, I decided to thoroughly investigate it, but only 13 occurrences were identified in the corpus. So, what follows represents just a tentative analysis that cannot be corroborated without referring to additional evidence. Despite this, I think it deserves to be examined exactly because of the reasons discussed above and, also, because it clearly shows the points made in the previous sections on the relationship between typology and discourse phenomena.

From the formal side, *ti'tuik prarö* construction is composed of two elements: a participial form of the verb 'know' (which is circumfixal, *t-V-ik*, like in proto-Cariban **te-V-ce* as reconstructed by Gildea 1998: 140-151) and a negative emphatic particle (6). The overall construction is literally translated as 'not knowing(ly)'.

(6) Morphological structure of *ti'tuik prarö*:

ti'tuik	***prarö***
t-i'tu-ze	*bra-rö*
ADV-know-PTCP	NEG-EM

'lit. not knowing(ly)'

[10] According to Glottolog 5.2, Akawaio is now considered a language and no longer a variety (the language is currently named Akawaio-Ingariko). However, since this contrast with the general view on Akawaio by experts and since this is not the objective of this paper to discuss the linguistic status of Akawaio, I still refer to it as a variety of Kapóng.

At the syntactic level, *ti'tuik prarö* can be found in four different positions in texts: (i) as a clause modifier (adverbial adjunct) (7a), (ii) as a V(P) modifier (7b), (iii) as a N(P) modifier (7c), and (iv) in autonomous position at the end of the clause (7d).

(7) Syntactic position of *ti'tuik prarö* (Caesar-Fox 2003: 317, 485, 310, 430):

a. ***t-i'tu-ze*** ***bra-rö*** *mïgï* *agïdï-bödï-bök* *y-eji*
ADV-know-PTCP NEG-EM this cut-PLAC-PROG 3-be
'**Unknowingly and without care**, he would cut up whatever he wants to'

b. *meguru yek* *pömï-u-ya* *chigaru* *yek* *pömï-u-ya*
banana plant plant-1-ERG sugar.cane plant plant-1-ERG
t-i'tu-ze ***bra-rö*** *Ø-e'-pömï* *ibïra* *rö*
ADV-know-PTCP NEG-EM 1-DETR-plant with.no.doubt EM
'I plant banana plants, I plant sugar cane plants, (or) I plant **anything else(/without knowing)**, no problem'

c. ***t-i'tu-ze*** ***bra-rö*** *murang* *eji* *mörö kwaro'nai* *kubi'ta*
ADV-know-PTCP NEG-EM charm be AI(?) ginger.charm herb.charm
kwak *ta* *tok* *ya*
charm(sp.) say 3PL ERG
'There are **numerous (types of)** charms, 'kwaro'nai, kubi'ta, kwak, so they say'

d. *ane* *ji* *an-egama-gö* *pandong* *wayamori* ***t-i'tu-ze*** ***bra-rö***
wait.IMP EM 3O.IMP-tell-IMP story turtle **ADV-know-PTCP NEG-EM**
'Please tell a story then, about the turtle **or anything**'

At the functional level, the situation is quite complex, mainly because of the scarcity of occurrences in the corpus. However, I was able to identify at least three different functions: (i) when *ti'tuik prarö* is used as a general extender[11] (see (7d) above); (ii) when *ti'tuik prarö* is used to express an heterogeneous set of entities (similar to an hypernym or a category label, see (7c) above); (iii) when *ti'tuik prarö* is used to express manner (see (7a) above). To these, I also identified a case in which the function is ambiguous, this is exemplified in (7b) where *ti'tuik prarö* can both have a GE or a manner reading.

[11] A general extender (GE) is generally defined as "a form that indicates additional members of a list, set, or category [and that combines] with a named exemplar (or exemplars)" (Overstreet 1999: 11), e.g., *and the like, and things like that, and so on, etcetera*, etc.

Despite the scarcity of occurrences, we can try to go further and propose a tentative path of evolution for the *ti'tuik prarö* construction (8) mainly based on the formal and functional properties just shown and on their frequency in the corpus, also in comparison with other constructions that are similar to *ti'tuik prarö*. Of course, this proposal must be conceived just as an attempt rather than an actual empirically-based diachronic scenario.

(8) Tentative path of evolution of *ti'tuik prarö*:[12]
Mod C (adverbial adjunct - manner) > Mod V(P) (adverb - manner, but also GE) > Mod N(P) (after a list of elements, heterogeneity or GE) > Autonomous (after a list, GE)

The first stage would probably be *ti'tuik prarö* as a clausal modifier encoding manner. This is, indeed, the original position and function of the **ti-V-ce* construction as reconstructed in Proto-Cariban by Gildea (1998: Ch. 8). I found only one occurrence (out of 13) of *ti'tuik prarö* used in such a way. I also checked other *ti-V-ze* constructions in the Akawaio corpus, and I found 12 occurrences (out of 27) of other participles with the same structure showing these function and syntactic position. The second stage would consist of *ti'tuik prarö* used as V(P) modifier (another original position of **ti-V-ce* construction in Proto-Cariban) expressing again manner, but eventually also employed as a GE. In this case, I found 2 occurrences (out of 13) in the corpus, and I also found 9 occurrences (out of 27) of other participles with *ti-V-ze* structure displaying this function/position. The third stage would predict *ti'tuik prarö* as N(P) modifier after a list of elements conveying what I called heterogeneity (i.e., a heterogeneous set of entities, hypernym/category label). The occurrences of *ti'tuik prarö* with this function/position are 4 in the corpus (out of 13). I found 5 occurrences (out of 27) of other participles with *ti-V-ze* structure that have this function, but these cases do not modify a N(P), strictly speaking, but a nominalized verb mainly with a predicative function (thus a much more verb-like entity). The fourth and final stage would suggest *ti'tuik prarö* in an autonomous syntactic position at the end of a clause (after a list) used as a GE. I found 6 occurrences (out of 13) of this case and no occurrences (out of 27) of other participles with the same structure of *ti'tuik prarö*

[12] This path of evolution was discussed by the author with Spike Gildea that I would like to thank for the generous support. However, all possible mistakes, misunderstandings, and misinterpretations (if any) must be considered solely mine.

showing these functional-syntactic properties. To sum up, it might be probable that *ti'tuik prarö* started being a truly participial form modifying clauses or V(P)s with an adverbial function (with some VP also like a sort of GE). These two situations are the less frequently found for *ti'tuik prarö* (3 out of 13) but the most common for other *ti-V-ce* constructions (19 out of 27). Then, it started being used as a modifier for N(P)s, maybe with situations employing nominalized verbs or nouns with a predicative function as a bridging context and then extended to "true" nominals (through reanalysis) expressing heterogeneity (because of its negative element). This situation is quite frequent for *ti'tuik prarö* (4 out of 13) but much less for other participial constructions (5 out of 27). Finally, *ti'tuik prarö* might have been reanalyzed as a GE, which is the most common situation for *ti'tuik prarö* (6 out of 13) and is not attested with other *ti-V-ce* constructions (none out of 27).

Of course, the evidence for such an evolution path is scanty and relies a lot on frequency in the corpus, which is ultimately based on very low figures, not allowing me to draw strong conclusions. However, the increasing frequency figures of *ti'tuik prarö* when going on the right along the path and the concomitant decreasing of occurrences of similar constructions are compelling evidence and, in my opinion, makes the path in (8) at least plausible.

In conclusion, we can say that Akawaio *ti'tuik prarö* is presumably developing as a particular discourse marker, i.e., a general extender. At the typological level, GEs are defined as elements with (Mauri & Sansò 2017: 65, my free translation):

> associative referential function […]. We can describe this function referring to three kinds of entities to which GEs make reference to:
>
> i. one or more explicit exemplars,
> ii. additional non-explicit elements X, that are associated to exemplars according to a shared property that is relevant for the context,
> iii. a wider category that includes both explicit exemplars and implicit additional elements X

We find all these properties in several of the situations in which *ti'tuik prarö* is found.

The *ti'tuik prarö* construction seems to be peculiar to Akawaio solely within the language family: I quickly consulted some corpora and grammars for other Cariban languages looking for similar constructions and I did not find any of them (also confirmed by Gildea p.c.). However, Trió (tri; Cariban, Guianan) displays a quite similar and interesting construction: *ookinenpen* (that.INAN-PST-CONT) 'all kinds of things (lit. those things that used to be)'. Even though this construction seems to have a

different function (rather heterogeneity – 'all kinds of X' – than a true GE), I only found 2 occurrences in the Trió texts and, in addition, this probably has a verbal origin, too, but not a cognate of *ti'tuik prarö*, leaving open possible speculations over its evolution path.

This very brief case study is particularly important for several reasons. The *ti'tuik prarö* construction represents an analytical strategy in which we cannot identify any connective, which is a typical property of GEs as described in the cross-linguistic literature[13] as proposed by Mauri & Sansò (2017: 66) for GE structure: connective + indefinite/generic element + similative element. In addition, *ti'tuik prarö* has a verbal origin that represents a totally new source for GEs cross-linguistically (cf. Mauri & Sansò 2017). At the same time, the pattern NEG + 'know' is not fully unknown for other types of discourse markers (e.g., *non so* 'don't know' in Italian, cf. Lo Baido 2020, or *I don't know* in English). Thus, this widens our understanding of GE in the languages of the world, showing how they might be more varied (both synchronically and diachronically) than attested so far, but also shows how there might be some commonalities between different discourse markers in different languages.

This case study might represent an important starting point. On the one hand, typologists must realize that giving the right importance to discourse phenomena and their cross-linguistic variety can provide important information, allowing us to account for possible evolution path involving different level of analysis (e.g., morphosyntax and discourse, like in the case of *ti'tuik prarö*) and thus better understand how languages shape their own grammar and why they are organized in such a way. Still, it would also allow typology to adopt a truly usage-based approach. On the other hand, discourse studies should be aware of and investigate the cross-linguistic variety (thus in non-WEIRD and non-LOL languages, too) of discourse phenomena since this can help in better describing, understanding, and assessing them also in single and specific WEIRD and LOL languages.

4. How can we make data speak to us: A multi-level methodology

The aim of this paper is not solely to discuss some issues and criticize the methods and the data on which typology has been founded so far. I would also like to propose

[13] However, we must bear in mind that in Cariban languages connectives are not very frequent strategies, these languages tend to employ more frequently juxtapositions.

a possible method that would help solve, at least partially, some of the issues considered in the previous sections.

We saw how the best kind of data for typology would be having "primary" data along with grammatical descriptions (corpora on which the latter are based on) since these could allow us to look for/at linguistic phenomena (from morphosyntax to discourse) in their own context/co-text. This is exactly what the abovementioned projects (e.g., MultiCAST, DoReCo, CorpAfroAs, etc.) have tried/are trying to do. However, as already noted, the path is very long, and this is not fully viable right now. So, what can we do? In my opinion, the only possibility we have is to follow a multi-level *converging evidence* method. The converging evidence method consists of adopting a multifaceted perspective to investigate the object of analysis, that is, taking data and pieces of evidence from different kinds of sources and perspectives. This is what Mauri & Masini (2022) call "the 3D methodology", which "combin[es] Discourse analysis with cross-linguistic Diversity and/or Diachrony" (Mauri & Masini 2022: 101) to which I would add sociolinguistic and areal/contact information (if available). In other words, for what concerns typology, this means to analyze linguistic phenomena from a cross-linguistic perspective, also looking at their diachrony (possible sources and evolution paths) and their discourse properties mainly through corpus-based language-specific analyses. All these perspectives are mutually connected, and comprising all of them is fundamental to maximizing the possibility to catch and account for the whole complexity of what we are investigating.

Typology, by definition, simplifies language complexity (found in single languages) in order to identify recurrent patterns and propose generalizations that can ultimately provide hints on the cognitive organization of information. However, "simplification" should not correspond to "oversimplification",[14] which is the most dangerous risk for typology. This is because oversimplified data would end up in possible not well-suited (or even wrong) cross-linguistic generalizations and predictions. In order to avoid oversimplifications, typologists need to base their investigations over the wider range of evidence as possible.

[14] I am aware that identifying a boundary between "simplification" and "oversimplification" is indeed difficult. However, discussing such an issue is not an objective of the present paper, and, in addition, it is not fundamental to my purposes. This is because the boundary is not discrete and may vary depending on the phenomenon and the research objectives. Typology should, by definition, simplify complexity to the bare minimum just to identify generalizations (regardless of where we put the boundary), and to do so, interpreting the data as best as possible is fundamental.

The method I am proposing here goes exactly in this direction and is based on the *converging evidence* perspective. This method (which can be simply called *discourse-sensitive typological method*) consists of a cross-linguistic 'multi-level method' that allows one to maximize the possibility of finding data and, thus, let emerge the widest cross-linguistic variety while looking at phenomena also in their discourse environment.[15] This multi-level typological method consists of three different levels of investigation: (i) the horizontal level (large-scale typological investigation), (ii) the intermediate level (a more 'qualitative' typological investigation), and (iii) the vertical level (intralinguistic investigations, i.e., case studies).

Horizontal level

The first level consists of a horizontal investigation, that is, the traditional way of doing typological research. The researcher will investigate the object phenomenon through a balanced sample of languages, generally a variety (and convenience) sample, composed of about 250 or more languages. Large variety samples are preferred over the other types because they are specifically designed to maximize the degree of cross-linguistic variety giving more relevance to internal complexity at the genealogical level and less importance to the actual statistical balancing (e.g., the Diversity Value – cf. Rijkhoff et al. 1993 and Rijkhoff & Bakker 1998 – and the Genus-Macroarea – cf. Dryer 1989, Miestamo 2005, Miestamo et al. 2016 – techniques). However, the sample size and type may vary according to the objectives of the research.

This level is based on grammar mining, as it is generally called within typology, which consists of looking for data through the analysis of grammatical descriptions. However, some potentially useful and practical techniques may be adopted to identify all possible patterns. For example, picking the "best" grammar for a language is very important. In general, the best choice would be the grammar that is most recent, the most exhaustive,[16] and the easiest to find. Another useful tool is to create a list of terms and glosses to which the object phenomenon can be referred to in the grammatical description in order to look for them (in the table of contents, the

[15] This method has already been tested in some works by the author of this paper. A preliminary description and an application of this method to a specific phenomenon can be found in Masini & Mattiola (2019).

[16] It is not easy to identify the most exhaustive grammar *a priori* (i.e., without reading it), so, in this case, the number of pages should be considered as an indicative clue.

analytical index, etc.) and thus detect all the possible information available. For example, for reduplication, we might use the following terms: *reduplication*, *repetition*, *duplication*, *multiplication*, *serialization*, *doubling*, *iteration*, *RED/RDP/REDUP*, and so on. This is particularly useful in digital versions of grammars that usually are automatically searchable (alongside other techniques of data mining).

Through this level, the researcher will have a first survey on the presence of linguistic phenomena and on how they work in a large sample of the world's languages. However, the general 'imperfections' of large-scale typology are still there, such as not having much information on the discourse usage and properties of the phenomenon. Despite this, the traditional typological method (which allows us only to scratch the surface of linguistic complexity) is still fundamental to have a general picture of what languages of the world display. However, as pointed out above, this method alone cannot suffice for a discourse-sensitive approach, and other levels of investigation are needed.

Intermediate level

The second level involves more detailed typological investigations. In this case, the researcher will design a smaller sample of languages (not necessarily balanced) of approximately 20-30 languages. Again, the numbers must be conceived as approximations and rather should adapt to the objectives of the research itself.

This level is based on a more fine-grained analysis of grammatical descriptions (e.g., taking in consideration different grammars or sources, like dictionaries, if relevant) and, alongside this, it also requires the analysis of texts of the languages, such as those found at the end of descriptive grammars or made available by linguists specifically working on that language (freely available or directly asking for them to experts). This phase is mandatory since it allows the researcher to look for and observe the phenomena directly in the texts of the language (also with the narrative context and linguistic co-text at one own disposal), and thus find also patterns that are not described within the grammatical descriptions (e.g., discourse phenomena) or not identified by the grammarian (e.g., because of the non-overlapping between descriptive categories – described within grammars – and comparative concepts – comparative definitions).

This level allows us to get over the problems of 'traditional' typology, pointed out in the previous sections, by verifying directly in a corpus (even if a small one) what the horizontal level might have not brought out. In addition, through the intermediate level, discourse starts playing a crucial role both in terms of description and in terms of explanation.

Vertical level

Finally, the vertical level consists of a much more detailed linguistic analysis of a very small language sample (from 2 up to 5 languages) comprising languages as typologically different as possible. This level relies on analyses of (large) corpora made available by linguists who are experts in a language (possibly first-hand data and glossed) or are freely available (e.g., on Sketch Engine). In this case, the researcher must be an expert in the language or should count on the help of experts since these are truly corpus-based analyses. Through these analyses, the full linguistic complexity of the phenomenon emerges (even if only for a few languages), making (virtually) all the existent patterns able to be identified and analyzed in detail. At this level, possible information on the discourse, pragmatic, and sociolinguistic levels might emerge more clearly, allowing for a thorough account of the object phenomenon.

This level is extremely important for typological analyses mainly because of two reasons: (i) it allows us to verify the typological generalizations identified in the first two levels and (ii) it allows us to investigate phenomena in much greater detail, letting possible characteristics and patterns (and eventually also unexpected parallelisms among different structures) that (traditional) typology cannot detect emerge.

The three levels must be considered as strictly intertwined and mutually dependent with each other. This means that if a pattern and/or a parameter of analysis, not previously considered, emerges from one of the levels, then the researcher should go back and forth through the levels and implement the analysis of each of them by integrating such a pattern and/or parameter accordingly.

This three-level method should help us in the difficult process of collecting consistent data in a typological sample of languages on linguistic phenomena (both “grammatical” and “discourse” phenomena) and analyzing them thoroughly. In this way, we retrieve the more varied data we can find in the sources we have at our disposal. This would help us a lot in proposing consistent analyses and strong generalizations. In fact, each level requires the adoption of the converging evidence perspective *per se* comprising cross-linguistic, diachronic, sociolinguistic, and areal data. All these data would “speak” to us and with each other, making our investigation and findings as solid and informative as possible.

In my opinion, a consistent methodology, like the one proposed here, is in order in (qualitative) typology. Even though some typological works have already introduced some of the phases of this multi-level method, it is still extremely difficult to find

investigations that fully comprise and adopt in a methodologically consistent way the suggestions made here.

5. Conclusion

This paper focused on some aspects of the scientific research (the data and the methods on which the discipline bases its research) on which typological community and literature have hardly discussed in detail. I first presented some theoretical preliminaries based on functionalist views that understand linguistic typology as a usage-based approach. I then focused on the aspects of such a perspective that typologists often tend not to fully integrate into their research. More specifically, I discussed two issues. First, I presented the different types of sources on which typological works are usually based, showing how they can generate possible issues for the analysis. Second, I discussed how typology tends to focus its attention on phenomena of specific levels (morphosyntax), almost completely disregarding phenomena of other levels of analysis (e.g., discourse). This represents an important shortcoming for typology since its final aim is to provide a comprehensive account and explanation to cross-linguistic variety of linguistic phenomena. Also in this case, I supported this view by exemplifying through a brief case study. Finally, I concluded the paper by proposing a multi-level discourse-sensitive method based on the converging evidence perspective that would allow (qualitative) typology and typologists to reach their objectives and finally become a truly usage-based discipline.

Acknowledgements

I would like to thank two anonymous reviewers for their valuable comments and suggestions, which helped me improve the paper. The ideas in this paper stem from personal reflections and challenges I faced during my research, as well as from discussions with people who shared their perspectives and expertise over the years. I would like to thank (in alphabetical order): Alessandra Barotto, Francesca Masini, Caterina Mauri, and Marianne Mithun. I also thank Silvia Ballarè for reading and commenting on a previous version of this work. Finally, I owe a special thanks to Spike Gildea, who generously shared the Akawaio texts, his expertise on Cariban languages, and several hours of his time to discuss the *-pödi* and *ti'tuik prarö* constructions with me. All shortcomings are my own.

Abbreviations

1 = 1st person
3 = 3rd person
ADV = adverb(ializer)
AI = addressee involvement
ATTR = attributive
CONT = continuative
DETR = detransitivizer
EM = emphatic
ERG = ergative
IMP = imperative
INAN = inanimate
NEG = negative
O = object
PL = plural
PLAC = pluractional
PRES = present
PROG = progressive
PST = past
PTCP = participial
STYLE = stylistic element

References

Barotto, Alessandra & Simone Mattiola. 2023a. Discourse phenomena in typological perspective: An overview. In Alessandra Barotto & Simone Mattiola (eds.), *Discourse phenomena in typological perspective*, 1-9. Amsterdam: John Benjamins.

Barotto, Alessandra & Simone Mattiola (eds.). 2023b. *Discourse phenomena in typological perspective*. Amsterdam: John Benjamins.

Becker, Laura & Matías Guzmán Naranjo. 2025. Replication & methodological robustness in typology. *Linguistic Typology* 29(3). 463–505.

Bickel, Balthasar. 2010. Capturing particulars and universals in clause linkage: A multivariate analysis. In Isabelle Bril (ed.), *Clause-hierarchy and clause-linking: The syntax and pragmatics interface*, 51–101. Amsterdam: Benjamins.

Bickel, Balthasar. 2011. Multivariate typology and field linguistics: A case study on detransitivization in Kiranti (Sino-Tibetan). In Peter K. Austin, Oliver Bond, David Nathan & Lutz Marten (eds.), *Proceedings of Conference on Language Documentation and Linguistic Theory 3*, 3–13. London: SOAS.

Bickel, Balthasar. 2015. Distributional Typology: Statistical inquiries into the dynamics of linguistic diversity. In Bernd Heine & Heiko Narrog (eds.), *The Oxford Handbook of Linguistic Analysis*, 901-923. Oxford: Oxford University Press.

Bybee, Joan L. 2006. From Usage to Grammar: The Mind's Response to Repetition. *Language* 82(4). 711-733.

Bybee, Joan L. & Clay Beckner. 2010. Usage-based theory. In Bernd Heine & Heiko Narrog (eds.), *The Oxford Handbook of Linguistic Analysis*, 827–855. Oxford: Oxford University Press.

Caesar-Fox, Desrey Clementine. 2003. *Zauro'nödok Agawayo Yau: variants of Akawaio spoken at Waramadong*. Doctoral dissertation, Rice University, Houston.

Chafe, Wallace. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In Charles N. Li (ed.), *Subject and Topic*, 25-55. New York, NY: Academic Press.

Chafe, Wallace. 1980. The deployment of consciousness in the production of a narrative. In Wallace Chafe (ed.), *The Pear Stories*, 9-50. Norwood, NJ: Ablex.

Chafe, Wallace. 1987. Cognitive constraints on information flow. In Russell Tomlin (ed.), *Coherence and grounding in discourse*, 21-51. Amsterdam: John Benjamins.

Chafe, Wallace. 1994. *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. Chicago, IL: The University of Chicago Press.

Corbett, Greville. No date. Agreement: Gender and Number. (Typological tools for field linguistics, MPI-EVA, Leipzig, available online at: https://www.eva.mpg.de/lingua/tools-at-lingboard/questionnaire/gender-and-number_description.php)

Cristofaro, Sonia. 2009. Grammatical categories and relations: Universality vs. language-specificity and construction-specificity. *Language and Linguistics Compass* 3(1). 441–479.

Croft, William. 2001. *Radical Construction Grammar. Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press.

Croft, William. 2003. *Typology and Universals, 2nd edn*. Cambridge: Cambridge University Press.

Croft, William. 2022. *Morphosyntax*. Cambridge: Cambridge University Press.

Croft, William & Alan D. Cruse. 2004. *Cognitive Linguistics*. Cambridge: Cambridge University Press.

Cysouw, Michael & Bernhard Wälchli. 2007. Parallel texts: using translational equivalents in linguistic typology. *STUF – Language Typology and Universals* 60(2). 95-99.

Dahl, Östen. 1985. *Tense and Aspect Systems*. Oxford: Blackwell.

Dahl, Östen. 2015. How WEIRD are WALS languages? Presentation at the conference *Diversity Linguistics: Retrospect and Prospect*, MPI for Evolutionary Anthropology, Leipzig, 1 May 2015.

Diessel, Holger. 2019. *The Grammar Network. How Linguistic Structure is Shaped by Language Use*. Cambridge: Cambridge University Press.

Dingemanse, Mark. 2012. Advances in the cross-linguistic study of ideophones. *Language and Linguistics Compass* 6(10). 654–672.

Dryer, Matthew S. 1989. Large linguistic areas and language sampling. *Studies in Language* 13(2). 257-292.

Dryer, Matthew S. 1997. Are grammatical relations universal? In Joan Bybee, John Haiman & Sandra A. Thompson (eds.), *Essays on Language Function and Language Type*, 115–143. Amsterdam: John Benjamins.

Dryer, Matthew S. & Martin Haspelmath (eds). 2013. *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at: http://wals.info (Accessed 2025.03.26).

Du Bois, John W. 1985. Competing motivations. In John Haiman (ed.), *Iconicity in syntax*, 343–365. Amsterdam: John Benjamins.

Du Bois, John W. 2003. Discourse and grammar. In Michael Tomasello (ed.), *The new psychology of language: Cognitive and functional approaches to language structure, Vol. 2*, 47–87. Mahwah, NJ: Lawrence Erlbaum Associates.

Gildea, Spike. 1998. *On reconstructing grammar: Comparative Cariban morphosyntax*. Oxford: Oxford University Press.

Givón, Talmy. 1979a. From discourse to syntax: Grammar as a processing strategy. In T. Givón (ed.), *Discourse and syntax*, 81–113. New York, NY: Academic Press.

Givón, Talmy. 1979b. *On understanding grammar*. New York, NY: Academic Press.

Givón, Talmy (ed.). 1983. *Topic continuity in discourse*. Amsterdam: John Benjamins.

Givón, Talmy. 1984. *Syntax*. Amsterdam: John Benjamins.

Guzmán Naranjo, Matías & Laura Becker. 2022. Statistical bias control in typology. *Linguistic Typology* 26(3). 605-670.

Haig, Geoffrey & Schnell, Stefan (eds.). 2023. *Multi-CAST: Multilingual corpus of annotated spoken texts. Version 2311*. Bamberg: University of Bamberg. Available online at: multicast.aspra.uni-bamberg.de (Accessed 2025.03.26)

Hammarström, Harald, Robert Forkel, Martin Haspelmath, Sebastian Bank. 2025. *Glottolog 5.2*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at: http://glottolog.org (Accessed 2025.12.09)

Haspelmath, Martin. 2007. Pre-established categories don't exist: Consequences for language description and typology. *Linguistic Typology* 11(1). 119–132.

Haspelmath, Martin. 2010. Comparative concepts and descriptive categories in cross-linguistic studies. *Language* 86(3). 663–687.

Henrich, Joseph, Steven J. Heine & Ara Norenzayan. 2010. The weirdest people in the world? *The Behavioral and brain sciences* 33(2-3). 61-83.

Hopper, Paul J. 1987. Emergent grammar. In Jon Aske, Natasha Beery, Laura Michaelis & Hana Filip (eds.), *Proceedings of the Thirteenth Annual Meeting of the Berkeley Linguistics Society: General Session and Parasession on Grammar and Cognition*, 139–157. Berkeley, CA: Berkeley Linguistics Society.

Hopper, Paul J. 1988. Emergent grammar and the a priori grammar postulate. In Deborah Tannen (ed.), *Linguistics in Context*, 117–134. Norwood, NJ: Ablex.

Lahaussois, Aimée & Yvonne Treis. 2019. Ideophones and interjections. Workshop at SLE 2019 Conference, University of Leipzig.

Lahaussois, Aimée & Julie Marsault and Yvonne Treis (eds.). 2024. Ideophones: honing in on a descriptive and typological concept. Special Issue of *Linguisdtic typology at the crossroads* 4(1). 1-444.

Majid, Asifa & Stephen C. Levinson. 2010. WEIRD languages have misled us, too. *The Behavioral and brain sciences* 33(2-3). 103.

Masini, Francesca & Simone Mattiola. 2019. Come fare tipologia con categorie non tradizionali? In Chiara Gianollo & Caterina Mauri (eds.), *CLUB Working papers in linguistics 3*, 282-294. Bologna: AMS Acta – Alma Mater Studiorum – Università di Bologna.

Mattiola, Simone. 2019. *Typology of pluractional constructions in the languages of the world*. Amsterdam: John Benjamins.

Mattiola, Simone & Spike Gildea. 2023. The pluractional marker *-pödï* of Akawaio (Cariban) and beyond. *International Journal of American Linguistics* 89(4). 457-491.

Mauri, Caterina & Francesca Masini 2022. Diversity, discourse, diachrony: A converging evidence methodology for grammar emergence. In Miriam Voghera (ed.), *From Speaking to Grammar*, 101-150. Berlin: Peter Lang.

Mauri, Caterina & Andrea Sansò. 2017. Un approccio tipologico ai general extenders. In Marina Chini & Pierluigi Cuzzolin (eds.), *Tipologia, Acquisizione, Grammaticalizzazione. Typology, Acquisition, Grammaticalization Studies*, 63-72. Milano: Franco Angeli.

Mettouchi, Amina, Dominique Caubet, Martine Vanhove, Mauro Tosco, Bernard Comrie, Shlomo Izre'el. 2010. CORPAFROAS, A Corpus for Spoken Afroasiatic Languages: Morphosyntactic and Prosodic analysis. In Frederick Mario Fales & Giulia Francesca Grassi (eds.), *CAMSEMUD 2007, Proceedings of the 13th Italian Meeting of Afro-Asiatic Linguistics*, 177-180. SARGON: Padova.

Mettouchi, Amina, Martine Vanhove & Dominique Caubet (eds). 2015. *Corpus-based Studies of Lesser-described Languages: The CorpAfroAs corpus of spoken AfroAsiatic languages.* John Benjamins: Amsterdam.

Miestamo, Matti. 2005. *Standard negation: The negation of declarative verbal main clauses in a typological perspective*. Berlin: Mouton de Gruyter.

Miestamo, Matti, Dik Bakker & Antti Arppe. 2016. Sampling for variety. *Linguistic Typology* 20(2). 233–296.

Mithun, Marianne (ed.). 1996. *Prosody, grammar, and discourse in Central Alaskan Yup'ik*. Santa Barbara, CA: Linguistics Department, University of California Santa Barbara.

Mithun, Marianne. 2015. Discourse and grammar. In Heidi Hamilton, Deborah Schiffrin & Deborah Tannen (eds.), *Handbook of Discourse Analysis. 2nd ed.*, 9-41. Oxford: Blackwell.

Overstreet, Maryann. 1999. *Whales, candlelight, and stuff like that: General extenders in English discourse*. Oxford: Oxford University Press.

Pakendorf, Brigitte & Françoise Rose (eds.). 2025. *Fillers: Hesitatives and placeholders*. Berlin: Language Science Press.

Ponsonnet, Maïa, Aimée Lahaussois & Yvonne Treis. 2023. Typologizing Interjections. Workshop organized at Dynamique Du Langage, Lyon, 21 november 2023.

Rijkhoff, Jan, Dik Bakker, Kees Hengeveld & Peter Kahrel. 1993. A method of language sampling. *Studies in Language* 17(1). 169–203.

Rijkhoff, Jan & Dik Bakker. 1998. Language sampling. *Linguistic Typology* 2(3). 263–314.

Schiffrin, Deborah. 1987. *Discourse Markers*. Cambridge: Cambridge University Press.

Schmidtke-Bode, Karsten, Natalia Levshina, Susanne Maria Michaelis & Seržant Ilja (eds.). 2018. *Explanation in typology: Diachronic sources, functional motivations and the nature of the evidence*. Berlin: Language Science Press.

Seifart, Frank, Ludger Paschen & Matthew Stave (eds.). 2024. Language Documentation Reference Corpus (DoReCo) 2.0. Lyon: Laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2).

Thompson, Sandra A. 1988. A discourse approach to the cross-linguistic category 'adjective'. In John Hawkins (ed.), *Explanations for language universals*, 167-185. London: Basil Blackwell.

Velupillai, Viveka. 2012. *Introduction to linguistic typology*. Amsterdam: John Benjamins.

CONTACT

simone.mattiola@unipv.it

Linguistic Typology at the Crossroads 5-2 (2025): 66-104 ISSN 2785-0943

Is complementation a universal strategy? A cross-linguistic corpus study*

NICHOLAS EVANS[1,2], DANIELLE BARTH[1,2], WAYAN ARKA[1,2,3], HENRIK BERGQVIST[4], CHRISTIAN DÖHLER[5], SONJA GIPPER[6], YUKINORI KIMOTO[7], DOMINIQUE KNUCHEL[8], DANIEL MAJCHRZAK[9], HITOMI ŌNO[10], EKA PRATIWI[11], SASKIA VAN PUTTEN[12], ANDREA C. SCHALLEY[13], ASAKO SHIOHARA[14], STEFAN SCHNELL[15], YANTI[16]

[1]SCHOOL OF CULTURE, HISTORY & LANGUAGE, AUSTRALIAN NATIONAL UNIVERSITY; [2]CENTRE OF EXCELLENCE FOR THE DYNAMICS OF LANGUAGE, AUSTRALIAN NATIONAL UNIVERSITY; [3]UNIVERSITAS UDAYANA; [4]DEPARTMENT OF PHILOSOPHY, LINGUISTICS, AND THEORY OF SCIENCE, UNIVERSITY OF GOTHENBURG; [5]BERLIN-BRANDENBURG ACADEMY OF SCIENCES AND HUMANITIES; [6]DEPARTMENT OF LINGUISTICS, UNIVERSITY OF COLOGNE; [7]GRADUATE SCHOOL OF HUMANITIES, OSAKA UNIVERSITY; [8]DEPARTMENT OF LINGUISTICS, UNIVERSITY OF BERN; [9]SCHOOL OF LITERATURE, LANGUAGES AND LINGUISTICS, AUSTRALIAN NATIONAL UNIVERSITY; [10]REITAKU UNIVERSITY; [11]FACULTY OF FOREIGN LANGUAGES, UNIVERSITAS MAHASARASWATI DENPASAR; [12]CENTRE FOR LANGUAGE STUDIES, RADBOUD UNIVERSITY NIJMEGEN; [13]DEPARTMENT OF LANGUAGE, LITERATURE AND INTERCULTURAL STUDIES, KARLSTAD UNIVERSITY; [14]RESEARCH INSTITUTE FOR LANGUAGES AND CULTURES OF ASIA AND AFRICA, TOKYO UNIVERSITY OF FOREIGN STUDIES; [15] INSTITUTE FOR THE INTERDISCIPLINARY STUDY OF LANGUAGE EVOLUTION, UNIVERSITY OF ZURICH; [16]ATMA JAYA CATHOLIC UNIVERSITY OF INDONESIA

Submitted: 06/02/2025 Revised version: 27/11/2025
Accepted: 27/11/2025 Published: 25/02/2026

Abstract

This article examines the question of whether complementation structures are cross-linguistically universal by using two different cross-linguistic corpora, each drawing on the same thirteen languages, spanning every continent. One is SCOPIC, the *Social Cognition Parallax Interview Corpus*, specifically designed to elicit material rich in grammatical categories relevant to social cognition; for each language in our sample this was balanced by

* Note (added on 3 March 2026): This article was updated to include Stefan Schnell in the list of authors. See the Corrigendum published at https://doi.org/10.60923/issn.2785-0943/24308.

a "general corpus" of roughly the same size with no specific targeting of domains. We find that, while complementation is widespread, it is not universal within the languages in our sample: in some it is absent entirely and in others it is extremely rare. Of the structural alternatives used to achieve the same functional goal by far the commonest is quoted speech, suggesting that in the evolution of linguistic structures it is heteroglossia, the embedding of one person's words in another's, that is a more basic phenomenon, from which complementation structures then evolve in many but not all languages.

Keywords: complementation; clause combining; corpus-based typology; clustering; parallax corpora; quoted speech; propositional framing

1. Introduction

In this article we ask two questions. The first is typological: are complementation constructions universal? This is a debate of long standing within typology (e.g. Aikhenvald & Dixon 2006[1]; Klamer 2000) but what is new about our study is that we subject the question to close cross-linguistic scrutiny using parallax corpus techniques. The second is methodological: does the SCOPIC corpus, a parallax corpus designed to elicit enriched naturalistic data on the language of social cognition (Barth & Evans 2017a, 2017b) provide a representative sample of language use more generally when compared against other datasets?

Recursion, of which complementation constructions are one of the three major types alongside relative clauses and adnominal recursion, is a cornerstone of much linguistic theorising, including accounts of generativity in grammar, vital steps in language evolution, and the development of *theory of mind* in social cognition, topics we summarise in (§2). Against this background, we need to ask a more basic empirical

[1] For example, in the abstract of their monograph on complementation Aikhenvald & Dixon (2006) write: "A **complement** clause is used instead of a noun phrase; for example one can say either I heard [the result] or I heard [that England beat France]. Languages differ in the grammatical properties of **complement** clauses and the types of verbs which take them. Some languages lack a **complement** clause construction but instead employ other construction types to achieve similar ends; these are called complementation strategies" (https://academic.oup.com/book/51146; accessed 2025-12-28), and their cross-linguistically inclusive collection gives examples of languages lacking complementation constructions. However, we believe that using the term *complementation strategies* continues to introduce a structural bias into the discussion, which is why we will use the more neutral term *propositional framing* as a functional label – see §3.3.

question: how widespread do complement constructions turn out to be, in fact, when we look at corpora of actual language use? Do all languages have them? Are they used more often for some types of complement than for others?[2] What are the functional alternatives to complement constructions? This is the clutch of typological questions we set out to answer in this paper.

In answering these questions, we draw on two distinct types of data (§3). One is the Family Problems picture task (San Roque et al. 2012) and the SCOPIC parallax corpus resulting from it (Barth & Evans 2017a, 2017b). Within the SCOPIC corpus (Barth & Evans 2024) we now have data for over 30 languages from every continent. In Kimoto et al. (2024), we drew on fourteen of these and examined what constructional alternatives to complement constructions are frequently attested in the SCOPIC data. However, it is important to ask how far it is representative of a broad sample of natural language use. In this paper, therefore, for each language represented, we also use a comparative "free" dataset – assembled from one or more corpora produced for other purposes; in §3.2 we outline how these free datasets, roughly comparable in size for each language to the corresponding SCOPIC sub-corpus, were built up across our language sample. Though there is substantial overlap in the languages used in the two studies, Kimoto et al. (2024) included Ilokano, Ku Waru, Jinghpaw, Japanese and Sibe, not in the present study, while this study includes Balinese, Avatime, Kogi, Komnzo, Vera'a and Yurakaré, not in the Kimoto et al. study. This reflected the availability of resources and time for the language-expert colleagues in our team to compile the "free" datasets needed for this part of the study.

In §3.3 we turn to the annotation schemata used in our analysis. Both sets of corpora were annotated for complement constructions and their functional equivalents, following an annotation scheme standardised across SCOPIC investigators; this uses a subset of the much larger set of conventions for annotating the SCOPIC corpus set out in Barth et al. (2024). These construction-type labels are applied to all propositional framing constructions in the corpus – basically all constructions which can frame propositions in the way that is done by complement constructions in at least some languages, as set out in the seminal article by Noonan (1985). We use this set of functional definitions to annotate structural alternatives to

[2] Again, we are by no means the first to ask this question. See e.g. Spronck & Casartelli (2021) for a wide-ranging survey of how complementation is likely to have originated through the intextualisation of direct speech; the extensive literature on how complementisers get recruited from 'say' verbs (Frajzyngier 1984, Klamer 2000, Saito 2021) implies that it is represented-speech complements which appear first in the grammaticalisation of complement constructions.

complementation in expressing comparable functions across the languages of our sample, in other words to examine the cross-linguistic semasiological possibilities for each relevant onomasiological choice. Thus, alongside complement constructions we also include such alternatives as paratactic quoted speech, inflections (e.g. desiderative inflections instead of 'want' verbs) and particles (e.g. counterfactual particles instead of verbs like *believe*). We likewise develop a set of functions (e.g. *utterance*, *thought*, *fear*) based on the functional range of complement-taking predicates in Noonan (1985).

In §4 we go back to our SCOPIC corpus and examine how far languages adopt the same coding strategies to express propositional framing. As we will show, the proportion of propositionally framed constructions that use complement constructions is highly variable cross-linguistically. While complement constructions are used commonly in some languages there are others in which they appear very rarely or not at all in our corpus, their place being taken by other constructions like paratactically linked direct quotation.

In §5 we evaluate the representativeness of the SCOPIC corpus by comparing the results for each language-specific SCOPIC sub-corpus with a comparably sized dataset from the same language. Overall, the results are remarkably similar, but there are nonetheless differences for some of the languages in our sample, which we discuss in this section.

In §6 we conclude by furnishing answers to the two questions posed at the outset. Taking these in reverse order, and beginning with the methodological question, there is broad agreement between our two types of data. Once we factor out specific reasons altering the occurrence of some strategies in the sub-corpora, we can take the SCOPIC corpus as an ecologically valid representation of overall frequencies of complementation or its functional equivalents. The great advantage of the SCOPIC data is that it permits close alignment of described scenarios across a wide range of languages while not biasing participant's expressive decisions through the effects of a "founder" language, such as one finds in parallel corpora (e.g. Mayer & Cysouw 2014), where one language is the source[3] and the others are translations from that source. In other words, it does not carry forward particular decisions about what to encode grammatically in an original "founder" language, from which translations are produced; rather speakers and signers of each language make their own direct decisions about what to encode based on the stimulus set.

[3] For the full Bible, of course, the source language is sometimes Biblical Hebrew (most of Old Testament), sometimes Aramaic (portions of Daniel and Ezra), and Greek for the New Testament. This does not alter the fact that, for whatever portion of the Bible we are talking about, one language is a "founder" or ultimate source for translations into all other languages.

An affirmative answer to the methodological question then gives us confidence to tackle the first, theoretical question. We find that complementation is not a universal encoding strategy, and some languages draw on other structures to achieve the same expressive ends, in particular by using direct quotation. The intertextuality afforded by direct quotation, we conclude, is a more cross-linguistically robust source of what may then, optionally and in only some historical trajectories, evolve into syntactically recursive complement constructions.

2. The claimed centrality of complementation to syntactic theory and social cognition

Over the last six decades, complementation has been central to arguments about the recursiveness of language, about the key steps in the evolution of language proper from earlier communicative systems and about the development of social cognition. Consider the multiple levels of complementation exhibited in (1):

(1) *All readers of this journal suspect that [Caterina thinks that [many typologists don't believe that [convincing typologies need naturally occurring discourse data]]]].*

The rewrite rules that sentences like (1) are believed to need (of the type (i) S → … NP; (ii) NP → S) were used by Chomsky (1957) and subsequent developments of generative grammar to argue for the indispensability of powerful transformational grammars necessary to account for the generativity of language. Of course, complement structures are just one type of syntactic recursion, alongside relative clauses and adnominal modification structures. Thus, we can represent the visual recursion shown in Figure 1 by a complementation structure, in (2), a recursive relative clause structure, as in (3), and the recursion of adnominal NPs, as in (4).

(2) *Chomsky and Halle remember that [back in the main photo they were thinking that [they were dressed rather formally in the river scene]]]].*

(3) *The photo [that shows a scene [where Chomsky and Halle hold a picture [that portrays them as young students]]] was taken by Michael Yoshitaka Erlewine.*

Figure 1. At Ling50@MIT, Morris Halle and Noam Chomsky holding a 1988 picture of them holding a picture of them in 1953[4].

(4) *We here reproduce [a slide of [an image of [a photo of [a scene of [Chomsky and Halle's youthful days]]]].*

Among these three types of syntactic recursion, complementation has a privileged status in important neighbouring fields, such as the psychology of social cognition. Psychologists of social cognition such as De Villiers (2000) and De Villiers & De Villiers (2003, 2014) argued that the acquisition of *theory of mind* and mastery of *false belief tasks* in developing children went hand-in-hand with the acquisition of complement constructions that allowed them to represent states of affairs as being held in the minds of particular social actors. De Villiers & Pyers (2002: 1057) discuss

[4] Photo and concept credit. The idea for this photo came from Sabine Iatridou (watch https://www.youtube.com/watch?v=csE-MsT_NN0 from around 1:40:40; accessed 2025.12.28) and it is discussed in the following blog by Kai von Fintel: https://www.kaivonfintel.org/morris-noam-recursion/ (accessed 2025.12.28) and the photo from Michael Yoshitaka ("Mitcho") Erlewine, with whose kind permission it is reproduced here. Our thanks to all three for their kind help and permission with this photo and story.

the claim that the development of complement structures in children is the best predictor of their performance on false-belief tasks: "[the] effect here is likely to be bi-directional, namely, understanding mental states undoubtedly makes it more probable and easier for the child to encode events in terms of people's beliefs, motives, intentions, emotions and so forth". The rationale for this argument is that mental attitude predicates, like those which are multiply embedded in (1), are the way that humans represent the mental states of others, and that this is done using conceptual analogues of the recursive complement structures shown here. And indeed, various language-specific modifications of complement structures are widespread cross-linguistically. For example, (5) from Japanese, produced spontaneously during one run of our SCOPIC task (§3.1), is an example with three levels of embedding of complement clauses. As is typical for a left-branching language like Japanese the embedded complement clauses each precede their complement-taking predicate.

(5) Japanese (Kazuya Inagaki - SocCog-jpn01-ikst3.eaf - 05:47.9-5:53.1)

de	*[[[de-ta*	*ato*	*wa*	*konna*	*koto*	*ga*
then	go.out-PST	after	TOP	such	thing	NOM

mat-teru	*daroo-na*	*to]*	*iu*	*no*	*o*
wait-PROG	may-FP	QUOT	say	COMP	ACC

soozoo-si-teiru]	*zu*	*da*	*to]*	*omoi-masu.*
guess-do-PROG	picture	COP	QUOT	think-POL

'Then, (I) think [(this) picture shows that [(he) is imagining [what will happen to (him) after going out (of prison)]]].'

Nonetheless, these structural solutions are not universal, and they are only used in a subset of the world's languages: languages have many other methods for showing mental attitudes. For example, in the Australian language Dalabon, the particle *djehneng* or its variant *yangdjehneng*, roughly 'believedly', is used to cast a statement as someone's belief rather than an actual fact; the identity of the believer is established pragmatically, e.g. through mention of a relevant belief-holding candidate in a preceding clause. Consider the following example from a text about a songman J's anger when he believes that another singer is singing J's compositions and passing them off as his own. Translation (a) puts this into more normal (complementiser-using) English, using the verb *think* and an embedded complement, while (b)

translates this more literally using an adverb to render the syntactic effect of *yangdjehneng:*

(6) Dalabon (Dal20090624.musdisc.mtandothers.eaf 2:58.4-3:01.8[5])

ka-h-kangurdinjirrmi-nj
3SG.SBJ-R-get.angry-PST.PFV

yangdjehneng *bûrra-h-marnû-dulu-djirdm-ey*
believedly 3DU.A > 3SG.OBJ-R-BEN-song-steal-PST.PFV

kanh *kodj-no* *barra-h-wayirni-nj*
DEM tune-3SG.POSSD 3DU.SBJ-R-sing-PST.IPFV

a. 'He got upset (that)/(because he thought that) the two of them had stolen it and were singing his song.'
b. 'He got upset; believedly they two had stolen it and were singing his song.'

Another Dalabon method for expressing mental attitudes without using complementation structures is shown in (7), this time using the adverbial prefix *molkkunh-* inside a polysynthetic verb.

(7) Dalabon (Evans et al. 2004: 245)
[Context: NE and a friend had turned up at the speaker's community the night before, without having been able to let her know, for want of a telephone, and camped nearby rather than imposing on her. Next day she reproached us:]

*de-h-**molkkunh**-bo-ng* *dabangh* *nahda,*
2DIS.SBJ-R-unbeknownst-go-PST.PFV yesterday hither

mak *yila-bengkey*.
NEG 1PL.SBJ > 1PL.OBJ.IRR-know-IRR

Lit. 'The two of you came here yesterday, unbeknownst, we didn't know.'
More English-like translation: 'We didn't know that the two of you had come here yesterday.'

[5] Available at https://www.gerlingo.com/language_detail.php?langID=7 (accessed 2025.12.18).

As with *yangdjehneng*, which in many ways is its semantic obverse (*yangdjehneng* specifying someone's belief, *molkkunh-* someone's ignorance), the clause containing it does not overtly specify the holder of the mental attitude. But as can be seen from these two examples, it is common (though not necessary) for this mental-attitude-holder to be specified in another clause nearby, such as *kahkangurdinjirrminj* 'he got angry' in (6) or *mak yilabengkey* 'we didn't know' in (7). In each of these two cases the identity of the mental-attitude-holder is not specified by the attitude-projection particle or affix itself (*yangdjehneng* 'believedly', *molkkunh-* 'unbeknownst') but found on another verb, whose subject identifies the holder of the mental attitude, in a paratactically juxtaposed clause. See Evans (2021) for many more examples, including cases where there is no neighbouring clause to specify the mental attitude-holder and this is left sheerly to pragmatic inference.

Here, then, Dalabon shows how it is quite possible for a language to represent mental predicates without the use of syntactic complement structures, in this case by a propositional-attitude particle showing projected belief. We will see later that it has other means as well, through the use of quoted speech; see also Kimoto et al. (2024) for various other constructional alternatives found in the SCOPIC corpora.

Within the field of language development, there have also been researchers citing other means of representing mental attitudes without the use of complementation strategies. For example, Matsui et al. (2009) in their comparison of theory-of-mind development in German and Japanese three-year-olds showed superior performance of Japanese as compared to German children on Theory of Mind tasks. They impute this to the high-frequency use of the illocutionary particle *yo* in Japanese, which signals a person's belief about what they are saying, as opposed to alternatives like *kana* 'maybe', and hence cues children early to attend to what those around them believe to be the case. Even though both Japanese and German have complement constructions in frequent use, for Japanese children it appears to be the illocutionary particles rather than the complement structures which are doing the heavy lifting in terms of scaffolding emergent social cognition, at least in the realm of theory of mind.

Considerations like these mean that, rather than assuming that complementation is cross-linguistically universal, and the sole structural means of representing mental attitude predicates, we should approach the question empirically and cross-linguistically. The approach we take is to define the relevant phenomena functionally and then determine what structure types are used to represent them using cross-linguistic corpora.

3. Datasets and corpora: methods and annotation

We now turn to the corpora for gathering comparable data across our cross-linguistic sample and the methods we use to annotate them: the *Social Cognition Parallax Interview Corpus* (SCOPIC) (§3.1), specifically developed to elicit enriched data on social cognition, and the less targeted general corpora (§3.2) which can be used to evaluate the representativeness or otherwise of the SCOPIC corpus, and the annotation schema we use in both (§3.3).

3.1 The cross-linguistic SCOPIC corpus

The cross-linguistic SCOPIC is a *parallax* corpus, which we have defined elsewhere (Barth & Evans 2017b: 1) as involving "broadly comparable formulations resulting from a comparable task", to avoid the implications of *parallel* corpus that there will be exact semantic equivalence across languages. The rationale for a parallax approach is that, by giving each participant the opportunity to respond in their own way to a shared stimulus, we leave it up to them to express things the way they want, without the "founder bias" that comes from strict translation tasks. In translation tasks, the source language Is likely to nudge the translation to transfer certain semantic categories or certain structures to the target language.

At the same time, using a shared stimulus gives us good control[6] over the referential characteristics of the described event, allowing us to match formulations across languages and across individuals.

The stimulus set consists of 16 picture cards from the Family Problems Picture Task (San Roque et al. 2012), which was developed as a *broad spectrum task* as part of a project on the cross-linguistic grammar of social cognition, with the goal of eliciting a wide range of themes relevant to social cognition (kinship relations, expression of emotions, private predicates, social consequences of actions from benefaction to malefaction), but also thought, speech, fear, memory and wishes for the future. These latter categories are all, evidently, good candidates for deploying complementation structures – but the overall task was not designed with any specific investigation of complementation in mind.

[6] Though of course this control is not complete, since different people construe the stimulus pictures according to their own cultural and individual schemata – clothes given back to a prisoner emerging from gaol may be seen as his own, or as a new ranger uniform; individual characters may be seen as male or female and so forth.

Our task was designed so that participants would see the task as meaningful and engaging, empathise with the characters and situations, including strong emotional reactions at certain points (such as domestic violence), understand the graphic conventions (speech and thought bubbles), understand the task specifications (breakdown into subtasks; ordering of pictures; distribution between dialogic and monologic subtasks) and discuss freely, vividly and unselfconsciously. We hoped to balance two aims: on the one hand to elicit broadly comparable cross-linguistic data, and on the other to be a kind of cultural Rorschach blot that calls forth interesting culturally specific and grammatically specific elements.

The task structure allows people, working in pairs, the chance to construct their own stories, across four task stages (a) initial card-by-card description, (b) joint construction of a meaningful narrative sequence (this stage was designed to elicit vigorous discussion), (c) narration of the story to a third party who had not been present for stages (a) and (b) and therefore started with no common ground, (d) narration of the same story, in the first person, from the perspective of one of the characters. See San Roque et al. (2012) for a detailed description of the task.

Over the last decade we have been building the cross-linguistic SCOPIC corpus from around 30 languages of all continents, including one sign language (Auslan); see Barth & Evans (2017b) for a listing of 25 of these, though the set continues to grow, and Barth & Evans (2024) for the archived corpus data including transcriptions and translations. In the research on which this article was based, we draw on a subset of thirteen languages (Table 1).

The sub-corpora for individual languages range from 34:01 (Indonesian) to 6:30:31 (Balinese) with a total of 33:39:49 hours and a mean of 1:40:59; for fuller details of SCOPIC sub-corpora sizes see §3.2.

3.2 Supplementary comparison corpora

As described in §3.1, the Family Problems Picture Task was designed as a broad spectrum stimulus task for getting enriched data on the expression of categories relevant to social cognition. SCOPIC data often includes expressions of mental attitudes, desire, intention, emotion, reported quotation, etc. This could potentially lead to corpus bias effects, e.g. if the choice of stimulus set were to bias the proportion of mental attitude constructions.

Language	Family	Location
Arta [atz]	Austronesian	Philippines
Avatime [avn]	Niger-Congo	Ghana
Balinese [ban]	Austronesian	Indonesia
Dalabon [ngk]	Australian	Australia
English [eng]	Indo-European (Germanic)	Australia
German [deu]	Indo-European (Germanic)	Germany
G|ui [gwj]	Khoe-Kwadi	Botswana
Indonesian [ind]	Austronesian	Indonesia
Kogi [kog]	Chibchan	Colombia
Komnzo [tci]	Yam	Papua New Guinea
Matukar Panau [mjk]	Austronesian	Papua New Guinea
Vera'a [vra]	Austronesian	Vanuatu
Yurakaré [yuz]	Isolate	Bolivia

Table 1: Languages used in the present study.

Additionally, the task uses certain visual devices (speech and thought bubbles appear in 7 of the 16 depicted scenes), which may prompt task participants to discuss speech and thought more than they would in other contexts. For the present study, therefore, we added a "supplementary sub-corpus" for each of the 13 languages and annotated that data according to the same schema (§3.3), to check the representativeness of our SCOPIC findings. As far as practicable the supplementary sub-corpus is (roughly) equivalent in size to the corresponding SCOPIC data for each language.

Whereas the SCOPIC sub-corpora are parallax, and thus broadly equivalent, the non-SCOPIC sub-corpora differ for each language, according to the contingencies of what language-specific investigators have gathered or have access to. Among the genres represented (summarised by language in Table 2) are: Pear stories (represented in 7/13 of the languages), Frog Stories,[7] traditional stories (including folktales), autobiographical narratives, conversation, sociolinguistic interviews and TV Debates. Table 2 summarises the amounts of data in each sub-corpus and labels the kinds of data found in the non-SCOPIC sub-corpora.

[7] An anonymous reviewer correctly observes that the Pear Stories and Frog Stories are also parallax. However, there is nonetheless an important difference in the relation of the resulting monologues to the stimulus. In both Pear and Frog Stories, the story line is already given by the stimulus, namely the order of episodes. In the SCOPIC task speakers were free to construct a storyline according to their own logic (and after discussion amongst themselves), meaning that choices of the order of framing (and hence of givenness, for example) are freer.

Sub-corpus	Length	PF Annotations	Annotation equivalency
Time-based sub-corpus information			***n* per minute**
Arta SCOPIC	1:22:55	274	3.3
Arta Pear Story, Traditional stories, Autobiographical narratives	1:16:29	141	1.84
Avatime SCOPIC	1:55:51	147	1.27
Avatime Traditional stories, Pear Stories	40:24	138	3.42
Balinese SCOPIC	6:30:31	745	1.91
Balinese Traditional stories, Spontaneous dialogue	1:33:03	229	2.46
Dalabon SCOPIC	1:20:40	420	5.21
Dalabon Traditional story, Autobiographical narrative, Pear Story (commentary and recall)	1:17:53	306	3.93
English SCOPIC[8]	3:02:50	569	3.11
English Sydney Speaks (sociolinguistic interviews)[9]	2:14:29	462	3.44
G|ui SCOPIC	56:00	214	3.82
G|ui Pear story, Interview, Traditional Stories	55:13	399	7.23
Indonesian SCOPIC	1:05:13	249	3.82
Indonesian Pear Stories, Autobiographical narratives, Traditional stories	34:01	195	5.73
Komnzo SCOPIC	1:20:59	160	1.98
Komnzo Narrative, Conversational Narratives	1:12:21	219	3.03
Matukar Panau SCOPIC	2:12:08	449	3.4
Matukar Panau Frog Stories, Exposition, Autobiographical narratives	2:13:37	127	0.95
Vera'a SCOPIC	1:13:44	129	1.75
Vera'a Pear Stories, Traditional and modern narratives, Local history	41:28	66	1.59
Yurakaré SCOPIC	1:36:02	513	5.33

[8] Five of our English SCOPIC sessions were collected by Gabrielle Hodge and Kazuki Sekine as part of a bilingual, multimodal corpus for comparison with Auslan (Hodge et al. 2019). Our thanks to them for providing this data.

[9] Travis et al. 2023. Our thanks to Catherine Travis for providing access to this data.

Sub-corpus	Length	PF Annotations	Annotation equivalency
Time-based sub-corpus information			***n* per minute**
Yurakaré Traditional stories, Sociolinguistic and other interviews[10]	1:41:30	245	2.41
Word count based sub-corpus information			**Per 1,000 words**
German SCOPIC	14,567	572	39.27
German Teacher feedback, TV debate, Narrative, Interview from Datenbank für Gesprochenes Deutsch[11]	21,193	1,037	48.93
Kogi SCOPIC	6,111 words	177	28.96
Kogi Frog Story, Pear Story, Traditional Story, Autobiographical narrative	2,145 words	80	37.3

Table 2: Sub-corpora summary information (PF = Propositional framing).

Including this supplementary comparison corpus allows us to determine if there is the same amount of propositional framing in SCOPIC vs other sub-corpora, and if it is of a different kind. This determination allows us to assess both the validity and reliability of using SCOPIC to answer questions relating to the typology of social cognition.
While SCOPIC and non-SCOPIC corpora are of roughly the same order of magnitude, for various reasons they are not identical. In general, if there is a discrepancy, the SCOPIC corpus is bigger, reflecting the fact that we have been annotating that over many years. These differences in sub-corpus size will be smoothed out by normalisation, as outlined in §4.

3.3 Annotation schema

All material from both the SCOPIC and the supplementary corpora were transcribed and translated using ELAN, a software for annotating audiovisual data developed at

[10] Data collection funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation, project 275274422, reference number GI 1110/1-1) Global South Studies Center, University of Cologne, DobeS initiative of the Volkswagen Foundation (grant numbers 81821 and 83448). Also includes data sourced from van Gijn et al. (2011).

[11] Selections from the *Datenbank für Gesprochenes Deutsch* (DGD): *Forschungs- und Lehrkorpus Gesprochenes Deutsch* (FOLK); *Deutsche Umgangssprachen: Pfeffer-Korpus* (PF); *Deutsche Mundarten: Zwirner-Korpus* (ZW), see IDS (n.d.).

The Language Archive, MPI Nijmegen, The Netherlands (e.g. Brugman & Russel 2004)[12] by the investigator(s) responsible within our team for that language, working with native language users if this was not their mother tongue. For each phenomenon of interest to our broader project we collectively developed an annotation schema, over a series of meetings of three continental "sub-teams" – Australian, Japanese and European – with Barth and Evans present at all of them and some other overlap of membership. This helped to support annotation reliability and typological categorical coverage across the sample of languages. Annotations cover a range of features relevant to social cognition, many irrelevant to the questions we are Investigating here, e.g. formulation of reference, use of kin terms; see Barth et al. (2021). For each of these features, values are annotated on a separate tier of our ELAN files. See Barth et al. (2024) and the appendix of Kimoto et al. (2024) for a full discussion of our annotation schema across all variables. Here we confine our discussion to the annotations relevant to complementation and the functions it expresses.

A fundamental question to resolve before we begin is whether complementation is defined structurally or functionally. Noonan (1985), in his classic typological overview of complementation, defines it as "the syntactic situation that arises when a notional sentence or predication is an argument of a predicate. For our purposes, a predication can be viewed as an argument of a predicate if it functions as the subject or object of that predicate" (Noonan 1985: 52).[13] But, after offering this syntactic definition, Noonan's chapter goes on to display some conceptual slippage between semantic and structural definitions. Thus on p. 64, picking up his earlier definition, he writes "in contrasting this *(universal) semantic characterization* [italics ours] with the surface characteristics of sentences containing complements…". Aikhenvald & Dixon (2006: 1), who maintain overtly that not all languages exhibit complementation structures, then propose the term *complementation strategies* for what happens in languages which, while lacking complementation structures, "still do have some grammatical mechanism for stating what a proposition is which is seen, heard, believed, known, liked, etc".

A fundamental assumption of typology is that we have to reckon with differential mappings of function onto form across languages. This creates a need to distinguish,

[12] Exceptions for supplementary corpora include Kogi and German data that were annotated based on textual transcriptions. Annotations were produced in either a text document or spreadsheet.

[13] Cf the definition in Aikhenvald & Dixon (2006: 4): "A complement clause has the following basic properties: (I) It has the internal constituent structure of a clause. (II) It functions as a core argument of a higher clause".

terminologically, between particular structures and the functions they express.[14] Now it often happens that there is terminological contamination between these two planes of linguistic analysis, as exemplified by the Noonan and Aikhenvald & Dixon quotes just given, but our preference is to keep the distinction as clear as possible by using entirely different terminology. We will therefore use the functional term *propositional framing* for the function and *complementation* for the structure as per Noonan's definition above.

The need for a comparable distinction between structure and function in the study of complementation has been repeatedly emphasised by scholars; in addition to the Noonan and Aikhenvald & Dixon passages above we can cite these lines from Deutscher's (2000) study of the evolution of complement structures in Akkadian:

> Different languages (or different diachronic stages of the same language) can use different structures to perform similar functions. For this reason, a valid comparison between languages needs to examine the role not only of complements, but also of other structures (such as parataxis), which can perform similar functions. This study differentiates between two concepts: 'complementation' (the embedding of a clause as an argument of a predicate), and the 'Functional Domain of Complementation', which includes complements as well as other strategies that perform similar functions (Deutscher 2000: 4).

Reflecting this need to distinguish claims about *structure* (here, complement structures) from those about *function* (e.g. expressing mental attitudes), our annotation scheme labels each relevant constructional token with a two-part annotation,[15] where the first part categorises the function and the second part its structure. Consider an annotation like UTT:COMP. This is to be understood as "utterance predicate [expressed by] complement structure"; it would be appropriate for an English sentence like *I told them that she arrived.* A complement structure could also be used for a probability judgment In English, e.g. *I guessed that she had arrived*

[14] Ideally, we would also have annotations for prosody, since there are languages like Teiwa (Sauerland et al. 2020) where complements appear to be paratactic if one confines oneself to morphosyntactic criteria, but can be argued to involve hypotactic structures once prosodic criteria are brought in. This would be an important extension of our investigations, but because comparisons across our set would only work if the prosodic analysis of each language in it was sufficiently advanced, we have been unable to integrate this into our annotations at this stage.

[15] This two-part schema, adopted for the present paper, is a simplified version of a more complex annotation schema used in other analyses by our team (Kimoto et al. 2024). In that schema, we also categorise the framing element (e.g. clause, particle), whether there is any connective present, and features of the content proposition (e.g. indirect speech). For current purposes we use a simplified version of this annotation, giving just the construction type and the functional domain.

or *It's likely that she has arrived*, and both would be annotated PROB:COMP. But an adverbial expression could also be used to express a comparable function, e.g. *She has probably arrived*, and this would then be annotated with the same function but with a different structure, as PROB:FUSE.

Our envelope of annotation possibilities, for propositional framing functions, takes in all functions which complementation structures express in at least some languages, as explored in Noonan's (1985) treatment. To assess the presence and distribution of propositional framing constructions across our data, we use the categories in Tables 3 and 4 respectively; for more detail see Kimoto et al. (2024) and Barth et al. (2024).

Abbrev	**Category**	**Example**
COM	Commentative	*I regretted that she had arrived.*
DES	Desiderative	*I hoped that she would arrive.*
FEAR	Fear	*I feared that she would arrive.*
IMM	Immediate perception	*I saw that she had arrived.*
KNO	Knowledge	*I knew that she had arrived.*
PRET	Pretence	*I pretended that she had arrived.*
PROB	Probability judgment	*I guessed that she had arrived.*
THINK	Thought	*I believed that she had arrived.*
UTT	Utterance	*I told them that she arrived.*

Table 3: Annotation categories for propositional framing functions.

Abbrev	**Structure**	**Example**
ADV	Adverbial clause (the framing element is in a subordinate clause to the clause expressing the content proposition)	*I am sorry, because the inspector has come.*
COMP	Complementation	*I regret that the inspector has come.*
COORD	Coordination	*I saw her and she was sleeping.*
FUSE	Proposition and frame fused into one clause: i.e., with a nominalised argument, verbal inflection or other bound morphemes, adverbs	*I regretted the inspector's arrival. / Maybe she is coming.*
INDP	Independent sentence with no framing element	*"She's come!"* [attributed speech uttered with no overt frame]
PARA	Parataxis	*I'm upset. The inspector has come.*
PRTH	Parenthetical	*The inspector, I can see, has arrived.*
SUB	Other subordination (including noun-modifying clauses)	*The picture of him arriving*

Table 4: Annotation categories for propositional framing structures.

Note that, to have a workable typology for comparison and to have sufficient tokens for statistical analysis, many of these categories are defined in a broad way. For example, FUSE includes bound adverbial prefixes (such as Dalabon *molkkunh-* in (7)) and other bound material such as desiderative inflections expressing 'want to', such as *-tai* in Japanese (*nometai* 'wants to drink'). While unsatisfactory for a maximally delicate typology, such borderline cases, and the functions they express, are relatively rare in our corpus, so that these simplifications do not affect our overall findings.

4. Cross-corpus comparisons

4.1 Quantitative Results: correlation heatmap dendrogram

In this section we first investigate complementation and its functional equivalents through quantitative means, across both corpus types, then use more qualitative methods to probe the reasons for the few differences we find between the SCOPIC and other sub-corpora. Our quantitative analysis is a means of seeing how similarly different languages pattern based on the ways they pair such structures as complementation, parataxis, fused constructions and other means of framing propositions with the various functions identified above. Our analysis also assesses how and whether the data type (SCOPIC or other) changes how languages pattern.

Our data included 6,973 annotations across all corpora. We excluded 12 tokens due to low use of the COORD structures. Other structural categories have between 136 (PRTH) and 2,702 (COMP) tokens. After exclusions, there are 6,961 tokens.

Figure 2 shows the distribution of construction types for each language, by each sub-corpus. Our take-aways from this figure are that the SCOPIC data often increases the amount of propositional framing used (see the taller bars for Balinese, English, Matukar, *inter alia*), and that a more heterogenous distribution of construction types tends to be found in the SCOPIC sub-corpora (see German [DEU] non-SCOPIC having primarily complement clauses, but SCOPIC data having also many fused-inflectional constructions and subordinating constructions). These are signs of task validity: namely that the task encourages people to produce more structures relating to social cognition (including propositional framing) than we would otherwise be able to observe in recorded data.

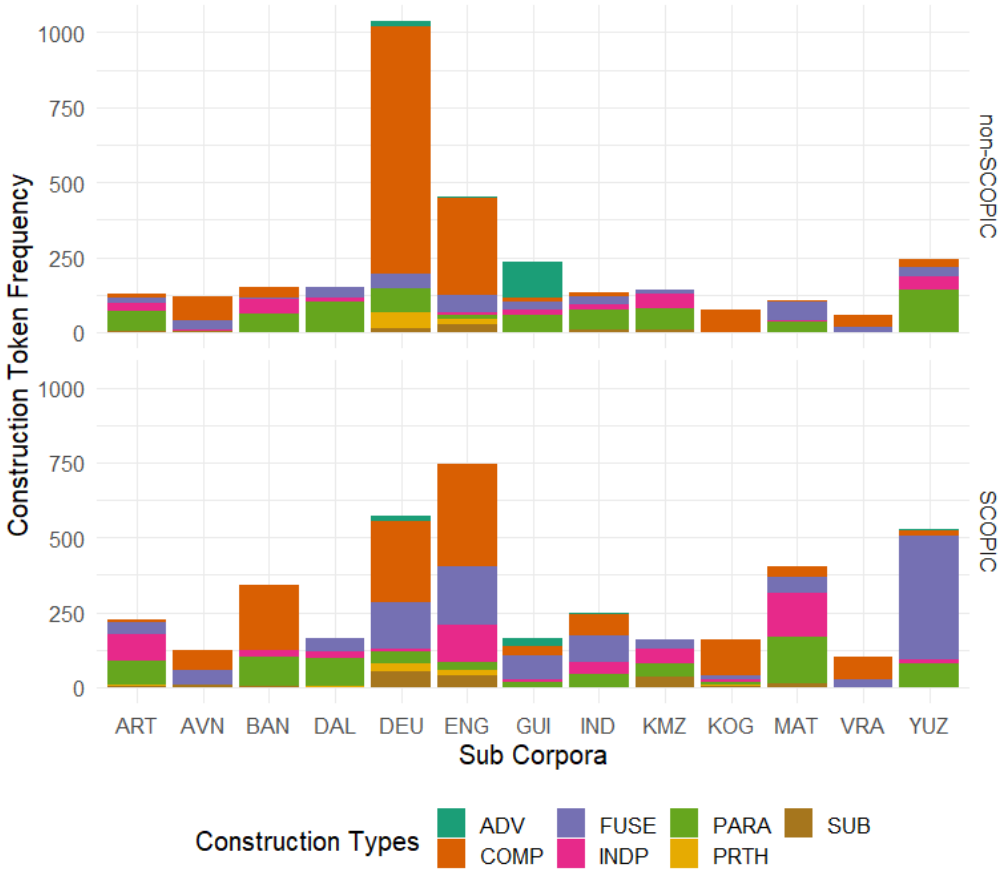


Figure 2: Construction types by language and sub-corpus.

Another important difference between the SCOPIC and non-SCOPIC data, particularly evident across quite a few languages (and most marked in Avatime, English, German, G|ui and Yurakaré) is the higher incidence of the FUSE structure in the SCOPIC sub-corpora (as shown by the increase in purple slices of the bar graphs). This reflects one important way in which the SCOPIC task does skew the data: during the SCOPIC task people use expressions of uncertainty that are integrated into the clause, like the adverbial expressions ENG *maybe,* DEU *vielleicht* 'possibly, maybe' when people are hazarding guesses about what particular cards depict and how they relate to each other (e.g. whether two cards contain the same character). This is reflected in Figure 3 which shows considerably more PROB tokens (light orange) for the SCOPIC data across many of the languages. However, we see similarities across the SCOPIC/non-SCOPIC data in that there often are a large amount of reported utterance tokens (dark orange THINK and purple UTT) and a small amount of pretence (red) and knowledge (pink) tokens.

Figure 3: Function types by language and sub-corpus.

Finally in our comparison of the bar charts, we notice that for some languages, there is a clear mapping between construction type and function type. Take Matukar, for instance, which has very similar type distributions in the construction and function types as UTT and THINK functions map directly to PARA and INDP constructions, DES functions map directly to FUSE constructions, and COM functions map directly to COMP constructions. This is not the case for our COMP construction heavy languages like German, English and Kogi, where the orange COMP slices of the bars in Figure 2 are very dominant, even when there is a more heterogenous distribution of function types. This is a sign that complementation is being used for a wide range of functions in these languages.

We now turn to the question of how similar and dissimilar the languages and sub-corpora are from each other based on their distribution of construction types. Investigating this question will help show whether complementation is predominant or not, and if not, what constructional alternatives there are available. Clustering analysis gives us insight into what kinds of typological usage profiles we find in the

different sub-corpora so we can consider its causes and the impact of the evolution of language complexity in terms of propositional framing.

We structure our data through "profiles of use". This means that for each sub-corpus we calculate a proportion of construction types used. We then evaluate how different each sub-corpus' profile of use is through a clustering analysis and heat map dendrogram created with the *R* package (R core Team 2023) *pheatmap* (Kolde 2019). We find four clear clusters in our data based on the proportion of use of each construction type. Figure 4 shows a heat map dendrogram with annotations labelling the clusters.

Heat map dendrograms are useful visualisations of differences as they combine two kinds of analyses: correlation heat maps and clustering dendrograms. The clustering dendrogram is a tree-structured representation of the data and how it clusters. Items (here sub-corpora) in each cluster are more similar to each other than to items from other clusters. This clustering is represented by the branches on the side and top of the heat map in Figure 4. Each branch shows which construction types (top dendrogram) and sub-corpora (side dendrogram) are best grouped together. Horizontal and vertical cuts of the data show where the strongest branching is. Lines connect the nodes that form clusters (King 2015). It is the clustering analysis that determines the order of constructions and sub-corpora along the x-axis and y-axis of these figures. The package *NbClust* (Charrad et al. 2014) was used to determine that the ideal number of clusters for the heatmap was four. Our analysis of this clustering leads us to characterise the clusters as [1] – COMP cluster, [2] - PARA/INDP cluster, [3] – FUSE/ADV cluster.

The correlation heat map uses colour to show how strongly entities in a matrix are associated. Each cell in our matrix represents how strongly a construction type and a sub-corpus are associated. The number in each cell reflects this association size. In our heat map in Figure 4, red is for "hot" to show a strong positive association between the sub-corpus and the construction type (positive value), and blue is for "cold" to show a strong negative association (negative value). Yellow shows no substantial association. As shown by labels on the x-axis in Figure 4, we are essentially counting the number of times each construction type was used within a sub-corpus to frame propositions. We then normalise these counts by centring and scaling the values (see Lucas et al. 2020). We group these measures by sub-corpus listed on the y-axis.

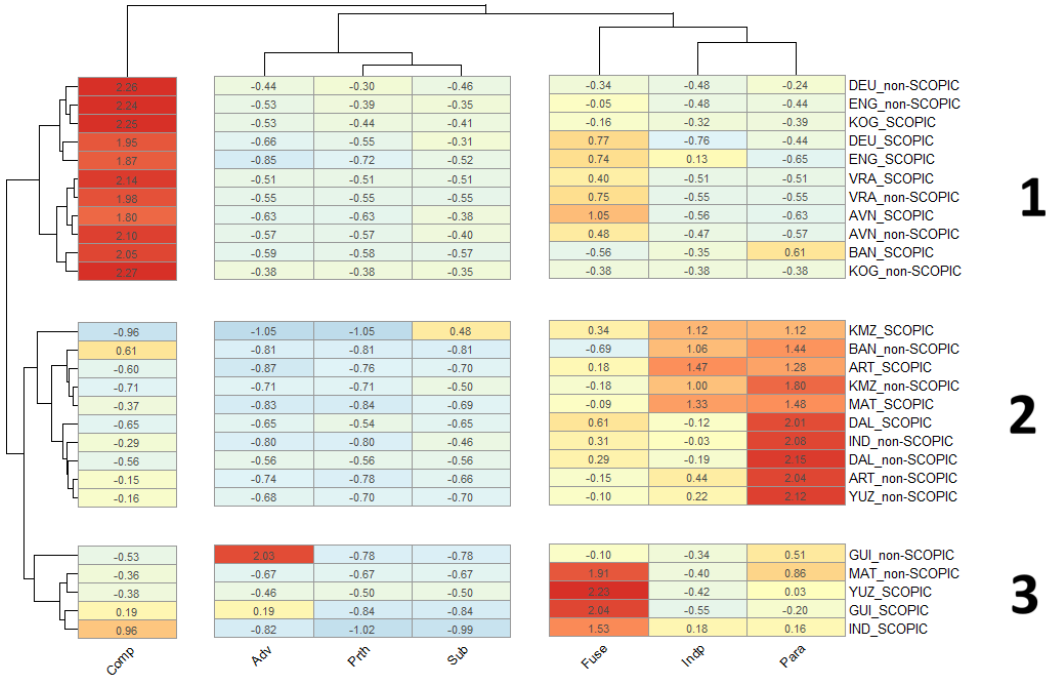


Figure 4: Heatmap dendrogram showing clusters of sub-corpora by the association with each construction type.

We now turn to the examination of each of the three clusters. More than half of the SCOPIC and non-SCOPIC sub-corpora for each language fall within the same cluster; we call these *sub-corpora pairs*.

Cluster 1 (COMP cluster) has sub-corpora pairs from German (DEU), English (ENG), Kogi (KOG), Vera'a (VRA) and Avatime (AVN), reflecting that these languages use high rates of complementation in all their data types. Balinese SCOPIC data is also in the COMP cluster. As opposed to the Balinese non-SCOPIC data, there are tokens from a wider range of functions present. These functions (DES, IMM, PROB) all map to COMP constructions, as well as about half of the reported utterances. This clustering reflects a pattern seen in the descriptive bar charts: if a language uses complementation structures, that structure is used for a wide variety of functions.

Cluster 2 (PARA/INDP cluster) has sub-corpora pairs for Dalabon (DAL), Komnzo (KMZ) and Arta (ART), reflecting a high amount of parataxis in all their data types. The cluster also has non-SCOPIC data from Yurakaré (YUZ), Indonesian (IND) and Balinese (BAN) and SCOPIC data from Matukar (MAT). For these four languages, there is a higher proportion of UTT and THINK functions than any other functions in their cluster 1 sub-corpora. This maps onto higher amounts of INDP and PARA constructions. Therefore, it is the function of reported utterance (expressed

paratactically in these languages) and its high presence in the sub-corpora that drives the clustering for the latter four languages.

Cluster 3 (ADV/FUSE cluster) shows three SCOPIC sub-corpora from Yurakaré and Indonesian. These sub-corpora showed a higher amount of PROB function tokens, which mostly mapped to the FUSE construction type by using adverbs to modify/frame propositional content. The Matukar (discussed further in section 4.2.1) non-SCOPIC sub-corpus is clustered in the FUSE group due to a higher proportion of DESID functions in that corpus, which map onto the FUSE construction type through the irrealis, desiderative inflectional suffix (i.e. proposition and frame within a single clause). There is a sub-corpus pair from G|ui in cluster 3, reflecting that neither COMP nor PARA structures are prevalent in the language. The G|ui SCOPIC sub-corpus, similar to Yurakaré and Indonesian, has a high amount of FUSE constructions from the PROB functions. The G|ui non-SCOPIC sub-corpus is the only sub-corpus with many ADV structures, these are due to that structure being mapped to THINK/UTT functions from a high expression of reported utterance in the G|ui traditional stories.

4.2 Qualitative Results: sub-corpora differences and similarities

As mentioned above, there are some cases where there are differences in construction types between the sub-corpora, within a single language. We have found through qualitative analysis that this reflects a difference in the functions expressed in the sub-corpora. Namely, some constructions do not regularly occur without expressing a particular meaning, and the difference emerges in our quantitative counts, as shown in Figure 4. We follow this below with a discussion of distributions in three languages: English (both sub- corpora in cluster 2 with high rates of complementation), Dalabon (both sub-corpora in cluster 1 with low rates of complementation) and Matukar Panau (data split across clusters 1 and 3).

4.2.1 Sub-corpora differences: Matukar Panau

Matukar Panau has few complement constructions, parentheticals, adverbials or other subordination strategies for propositional framing. While Matukar Panau in general has considerable subordination, at least on the not uncontroversial assumption that clause chaining and serial verb constructions are types of subordination, these are not used for layering ideas. Rather, the subordination is used to show temporal links

between actions (clause chaining like ‘go, see, then stay’), or composite events (serial verb constructions like ‘meet-eat’ meaning ‘feast’). Propositional framing subordination is fairly rare no matter the sub-corpora of Matukar Panau. To frame or comment on ideas, Matukar Panau uses paratactic and independent constructions, especially in the SCOPIC data. In the non-SCOPIC data, we see considerable inflectional construction usage in the frog stories, exposition and autobiographical narratives. We find that this is directly related to the semantic types used in the sub-corpora.

A common functional type across the sub-corpora is the desiderative. The sole FUSE propositional framing usage is for desideratives through an inflectional verbal suffix (as in Example 8). Additionally, in both sub-corpora, quoted speech and thought are expressed through independent and paratactic constructions. The reason the SCOPIC data is in a different cluster in our clustering analysis is that there is simply a lot more quotation in the SCOPIC texts (as in Example 9), reflecting their different profile of use. Therefore, the difference is not in the kind of syntax, but in the frequency of its usage. The distributions of the main kinds of constructions and their associated functions are laid out in Table 5. Other construction types are too infrequent to be worthwhile including here. Matukar Panau does have a (quasi-)complement clause structure, but it is infrequent and is restricted to commentative functions with fully inflected verbal clauses framed by non-verbal predication as in (10).

Data	Desiderative	Quoted thought	Quoted speech
SCOPIC Para		51	87
SCOPIC Indp		73	79
SCOPIC Fused	41	1	
Frog Story Para			1
Frog Story Indp			
Frog Story Fused	6		
Autobio Narrative Para		3	26
Autobio Narrative Indp		7	8
Autobio Narrative Fused	15		
Exposition Para			1
Exposition Indp			
Exposition Fused	39		

Table 5: Primary Matukar Panau Construction-Function distribution by text type.

(8) Matukar Panau (Tomas Taleo Kreno – DGB1-2013_031-Tomas_Taleu_Kreno_Frog_Story – 1:02.5-1:09)

gaun	*bab-bai*	*bul-do*	*aim*	*main*	*tuli-nge*
dog	bark-I.IRR.DESID	try-R.D	boy	TOP	say-I.R.PFV

“awa-m	*tau!”*
mouth-2SG	shut

‘The dog wanted to start barking and the boy said “shut up!”’

(9) Matukar Panau (Tomas Taleo Kreno – SocCog-mjk01-tk_jb_1 – 5:18.6-5:23.6)

garma-n	*alo=a*	*suse-nge*	*so-nge*
head-3SG	back=LOC	thread-D.SEQ	come-R.D

so-nge	*tamat*	*numa-n=te*
come-D.SEQ	man	hand-3SG=LOC

“so-ndop	*ngau=da*	*t-a”*	*i*	*bal-e*
come-D.IRR	1SG=COM	1PL.EXCL.SBJ-go	3	say-I.R.PFV

‘He tied his hair back and came, the man took her hand and he said “come, let’s go (lit: with me, we go)!”’

(10) Matukar Panau (Mingkui Agid – SocCog-mjk10-ckd_ma_2 – 4:39.1-4:40.6)

hum-e	*main*	*uyan*	*ti*
hit.PL-I.R.PFV	COMP	good	NEG

‘His hitting them is not good’

4.2.2 Sub-corpora similarities: English

English (Table 6) has a good number of complement clauses, and they are used consistently across a range of functions. The comparative corpus is the Sydney Speaks corpus (Travis et al. 2023) and consists of sociological interviews. The most frequent functions in both text types were knowledge, probability, quoted thought and quoted speech. We also see the same pattern of construction types across the semantic functions in both text types: complementation for all functions, fused for probability

and some independent (unframed) quoted speech. Other combinations of constructions and functions are infrequent and not different enough to cause the text types to appear in different clusters in our clustering analysis.

Data	Knowledge	Probability	Quoted thought	Quoted speech
SCOPIC Indp		2	1	57
SCOPIC Fused		166	1	10
SCOPIC Complement	14	110	35	140
Sydney Speaks Indp				9
Sydney Speaks Fused		56		
Sydney Speaks Complement	26	76	47	53

Table 6: Primary English Construction-Function distribution by text type.

Where we do see some differences is in the specific form of complementation for quoted speech. In our SCOPIC data, the framed quote is often indirect rather than direct. In the Sydney Speaks data, it is more likely to be direct. Further, in the Sydney Speaks data, the connective between clauses is more often *like* than *that*. Additionally, in our SCOPIC data the quotee is always a specific and identifiable referent (example (11)). In the Sydney Speaks data, we observe that the quotee in the framing clause is sometimes a non-specific/impersonal referent as in ***it*** *was like* or ***stuff*** *was like* (as in example (12)). This is an indication that complementation/quotation in this construction is still an area that is undergoing grammaticalization, since the Sydney Speaks interviewees in our data are younger than the SCOPIC participants[16].

(11) English (Kat – SocCog-eng06-BK_English_Story_Oct3118pm – 24:25-24:29.7)
And he's like "Julie you've been fooling in with the shop assistant"

(12) English (SydS_AYF – SydS_AYF_128 – 9:48.5-9:53.2)
it can't just go straight to her because things are piling up and ***stuff was like "wa!"***

[16] Note that all of our English speaking participants are from Australia and speak an Australian variety of English. Therefore this more likely to be a generational change than a dialect difference.

4.2.3 Sub-corpora similarities: Dalabon

When we compare the texts making up the non-SCOPIC Dalabon sub-corpus we see a very clear pattern: by far the most common propositional framing use in any sub-corpus is paratactic quoted speech (13). Additionally, we see some fused constructions for probability in almost all of the texts, in the form of the particle *kardû* 'maybe' (14, 16, both from the Pear Story). Quoted thought is not common outside the SCOPIC sub-corpus, though the boundaries between quoted speech and quoted thought-as-speech can be hard to draw (15, 16)[17] but in the SCOPIC data and the Pear Story data from the non-SCOPIC corpus it follows similar distributions as quoted speech. Figures illustrating the comparison between the SCOPIC and non-SCOPIC data for Dalabon are given in Table 7.

(13) Dalabon (MP Pear Story Recall 0.2.43-0.2.48.5)

ka-h-na-ng	*"Ngale!*	*kardû*
3SG.SBJ > 3SG.OBJ-R-see-PST.PFV	hey	maybe

nga-h-dja-lng-karrû-bakm-inj"
1SG.SBJ-R-just-SEQ-leg-break- PST.PFV

'He looked (and said/thought/saw): "Hey, maybe I've broken my leg just now."' (Alt. transl.: 'He saw that maybe he had broken his leg.')

(14) Dalabon (MP Pear Story Commentary 01.31-01.33)

kardû	*ba-ru-rr-inj*
maybe	3SG.SBJ.PST-rub-RR-PST.PFV

'Maybe he rubbed himself.' (i.e. rubbed his leg because he fell).

(15) Dalabon (MP Pear Story Commentary 01.33-01.35)

"ngaleh!	*nga-h...*	*nga-h-dordm-iyan!"*	*ka-h-yin-inj*
hey	1SG.SBJ-R	1SG.SBJ-R-go.up-FUT	3SG.SBJ-R-say/do-PST.PFV

'"Hey, I'll climb up" he said/went/thought.'

[17] Cf Reesink (1993) who discusses this issue in Papuan languages.

(16) Dalabon (MP Pear Story Commentary 3.57-04.04)

kah…	*kardû,*	*wonarr-inj*	*“ngale*	*kardu*	*nga-h-lng-darrû…*
3SG.SBJ	maybe	think-PST.PFV	hey	maybe	1SG.SBJ-R-SEQ-leg

darru-bakm-inj	*bah*	*kahke*	*ka-h-dja-mon*
leg-break-PST.PFV	but	nothing	3SG.SBJ-R-just-good

‘Maybe he thinks “hey! maybe I’ve broken my… my leg, but nothing, it’s OK”’.

There are interconnected issues of lexicography, translation and corpus analysis here, well illustrated by (16). The stem *wonarr-* ‘think’ is formally the reflexive form of *wona* ‘hear’, so ‘to hear oneself’, a common pattern in Australian languages (Evans & Wilkins 2000: 571, which contains another and contextually comparable use of the same verb). So, one more literal translation of (16) would be ‘Maybe he heard himself saying: “Hey! Maybe I’ve broken my leg”’ while a less literal one, taking the semantic transition to ‘think’ into account, is ‘Maybe he thinks “Hey! Maybe I’ve broken my leg.”’ Examples like this illustrate the difficulty of reaching an analytic decision about whether to treat them as quoted speech or quoted thought, but at the same time show how natural it is to adopt paratactic quoted-speech constructions for representing mental states.

Data	Probability	Quoted thought	Quoted speech
SCOPIC Para		47	272
SCOPIC Indp		5	34
SCOPIC Fused	44	6	
Pear Story Para		3	22
Pear Story Indp			11
Pear Story Fused	1		
Autobio Narrative Para		1	63
Autobio Narrative Indp			6
Autobio Narrative Fused	20	2	
Event Recall Para			5
Event Recall Indp			
Event Recall Fused			
Traditional Story Para			138
Traditional Story Indp			3
Traditional Story Fused	17	1	

Table 7: Primary Dalabon Construction-Function distribution by text type.

5. Conclusions and implications

To our opening question, regarding whether complementation constructions are universal, our study suggests a negative answer. They are certainly common, occurring as the dominant structural type for both corpora in close to half the languages of our sample – Avatime, English, German, Kogi and Vera'a – and as the dominant type in the non-SCOPIC corpus in one more (Balinese). And they are attested, at least once in both corpora, in all languages except Dalabon. On the other hand, they are entirely absent for one language in our corpus (Dalabon; both corpus types[18]), entirely absent from the non-SCOPIC corpus in Komnzo, and only occur at low frequencies in both corpora for Arta, G|ui, Indonesian, Matukar Panau and Yurakaré, and low frequencies in the non-SCOPIC corpus in Balinese. Paratactic structures dominate in Dalabon and Arta, as well as the non-SCOPIC corpora for Indonesian, Komnzo and Yurakaré. For G|ui, other subordination strategies predominate, and fused strategies are dominant in for several languages in the SCOPIC corpus (English, German, Indonesian, Yurakaré). As discussed in §4, however, the elevated occurrence of fused strategies in these latter languages reflects the large number of words like *maybe* elicited by people qualifying their speculative interpretations of pictures in the task.

Our finding regarding the non-universality of complementation fits in with a number of claims by other scholars. For the oldest varieties of Akkadian, spoken around 4,500 years ago, Deutscher (2000) argues that complementation was absent and that it was only later that the erstwhile causal subordinator *kīma* gradually developed into a complementiser, via a reanalysis path from causal adverbial clause structures of the type *He said/spoke to the governor* ***because*** *(k-marker) the barley was not collected* to complement structure of the type *He said/spoke to the governor* ***that*** *(k-marker) the barley was not collected*. Givón (1991) proposes a somewhat similar

[18] Our assertion about the lack of complementation in Dalabon is based on the two corpora in our study. A more far-reaching survey of Dalabon grammar turns up one highly specialised construction that could be analysed as complementation: 'want' verbs with different or partially disjoint subjects, which are clearly conventionalised biclausal constructions where the first 'want' verb is marked with a benefactive applicative which uses indirect object marking to index the subject of the complement. See Evans (2006, 2021). These constructions are extremely marginal, as indicated by their complete absence from the two corpora reported on here and their virtual confinement to elicited settings. In another Dalabon corpus of around 60 hours (Ponsonnet 2013), Maia Ponsonnet (emails to NE, 7/1/2025 and 10/1/2025) found just three examples.

scenario for the Biblical Hebrew complementiser *kī*, but this time revolving around framing predicates like *be happy / regret*; again a reanalysis from *be happy because X* to *be happy that X* is a small step semantically, providing a clear bridging context for the structural reanalysis from adverbial clause to complement clause.

More recently Hernáiz Gomez (2024), drawing on a larger corpus of the earliest forms of Akkadian, disputes Deutscher's claim that complement structures were entirely absent, but nonetheless goes on to show that in a number of Semitic languages with lengthy written traditions what are now clear complement structures originated as similative manner expressions.

These arguments gel with diachronic studies from elsewhere in the world that show how complement structures can emerge, by such means as the grammaticalisation of report verbs into quote markers and complementisers in the Austronesian languages Tukang Besi and Buru (Klamer 2000), the progression from paratactic to hypotactic structures (Harris & Campbell 1995), the tightening up of intonational links from *unbound* to *bound* (Diessel & Hetterle 2006) and the reanalysis of clausal or intonational boundaries so that *'He said this/that. "X"'* becomes *'He said [that X]'*, as shown for Mohawk by Mithun (2025).

Regarding recursive structures more generally, Widmer et al. (2017: 799), in their careful diachronic study of Indo-European, have shown, for recursive NP embedding, that "every type of NP embedding – genitives, adjectivisers, adpositions, head marking, or juxtaposition – is unavailable for syntactic recursion in at least one attested language. In addition, attested pathways of change show that NP types that allow recursion can emerge and disappear in less than 1,000 years". The net effect of these studies is to show that complementation, like other types of recursion, is by no means a universal structure: rather, it is something that evolves and sometimes disappears over time.

The reader will have noticed that, in our discussion of how complement structures emerge from various other structural sources (causal adverbs in Ancient Akkadian, speech reports in the cases examined by Klamer and Mithun), the material corresponding to the complement in a language like English is treated as reported quotation. Revealingly, this is why Dalabon, the most complement-averse language in our sample, makes such extensive use of parataxis: direct speech, whether actual or represented/projected, is combined with a wide range of framing verbs to convey the equivalent of complement structures in a language like English; in our corpus

these include such verbs as *bengdinj* 'was thinking', *yolhwehmun* 'feels bad, worries', *bengkang* 'thought', *kurnh-bengkabengkang* 'thought, worried', as well as other nominal framing devices like *men-no* 'his/her mind' (see further discussion and examples in Rumsey et al. 2022). This suggests that the most primal origins for such structures are to be found, not in syntax, but in potentially recursive embeddings of passages of quotes within one another, in other words intertextual embedding at the narrative or discourse rather than the syntactic level. This is the argument advanced in an important recent article by Spronck & Casartelli (2021: 19):

> The type of linguistic structures specifically dedicated to this task are reported speech. If linguistic reflexivity, that is, thinking and talking about language, is at the heart of the complexification of grammar, reported speech is at the heart of language evolution, which would at once explain its universality in the languages of the world and its relation to grammatical categories.

The present study, by pinpointing the many constructional alternatives to complementation that exist across a parallax corpus, helps clarify why it is not a necessary structure, since languages can employ many other means to realise the same communicative goal.

Grzech & Bergqvist (2025), in their introduction to a volume on the typology of evidentials and epistemics, highlight the growing realisation within linguistics that to truly understand language, communication and cognition, we must look "beyond single minds toward cognition as a process involving interacting minds" (Dingemanse et al. 2023: 1), and our need for methods that allow us to do this in a systematic cross-linguistic mode. They particularly mention the importance of Corpus Based Typology (Schnell & Schiborr 2022, Levshina 2022) as a way of capturing "intra-linguistic variation in language use and its relation to aspects of language systems" (Schnell et al. 2021: 6), and the need to ensure that such corpora, of which SCOPIC is an example, "are interactive and purposefully designed to explore social cognition, and allow an insight into how knowledge rights and obligations are negotiated in dialogic interaction" (Grzech & Bergqvist 2025: 15). In this article we have argued that appropriately designed corpora can indeed furnish information relevant to social cognition and can contain sufficiently high occurrence rates in domains of interest to reach statistically robust conclusions. At the same time, these results are compatible with the patterning found in less targeted corpora.

More specifically, using such corpora to examine the cross-linguistic occurrence of complementation or its functional equivalents allows us to see that complementation structures, though common, are not universal and that there exist a significant number of structural alternatives to them, most importantly paratactic constructions involving represented speech (Kimoto et al. 2024). These paratactic constructions are employed not just for speech per se, but for "internal mono/dialogue" accompanying thoughts, memories, intentions and perceptions. By showing that it is these structures, rather than syntactically specialised complementation constructions, which are truly ubiquitous, we are drawn back to the Bakhtinian insight that it is *raznorečie* or heteroglossia, the threading together of different people's words, which is what is truly universal in how we construct infinitely large and complex linguistic units from finite numbers of words and construction types. At the same time, the common grammaticalisation pathway by which complementisers can evolve from elements introducing quoted passages, whether verbs (e.g. Klamer 2000 on Austronesian), demonstratives (Mithun 2025 on Mohawk), similative/manner expressions (Hernáiz Gomez 2024 on Semitic) or causal subordinators (Deutscher 2000 on Akkadian), indicates how it is possible for the widespread occurrence of complementation constructions across languages to be linked to its roots in quoted speech.

Acknowledgements

This work was funded by the Australian Research Council (ARC; DP0878126, DP130101655, DP140102124, FL130100111), the *Anneliese-Maier Forschungspreis*, awarded to Evans by the Alexander von Humboldt Foundation and the German Federal Ministry of Education and Research, the *ARC Centre of Excellence for the Dynamics of Language (CoEDL)* (ARC; CE140100041), the *Deutsche Forschungsgemeinschaft* (DFG; 275274422, 417675039), the *Japan Society for the Promotion of Science* (Kakenhi grants JP18K00582 and JP22K00536), the *Marcus and Amalia Wallenbergs Minnesfond* (MAW; 2017.0081), the *Swedish Research Council* (VR; 2017-01969 and VR; 2020-01581), the *Swiss National Science Foundation* (P0BEP1_165335), the *UK Arts and Humanities Research Council* (AHRC; AH/N00924X/1), the Global South Studies Center, University of Cologne, and the Volkswagen Foundation DoBeS program (VW: 81821 83448, 85606, 86357). We

thank the above-named institutions for their generous support of our research. We also thank Silvia Ballarè, Simone Mattiola, Nicola Grandi & Caterina Mauri, convenors of the conference at which these ideas were presented, *Naturally occurring data in and beyond linguistic typology* (Bologna, 18-19 May 2023), for their kind invitation to present there, two anonymous referees for their useful critical comments on an earlier version of this paper, Keira Mullan for assistance with the formatting, and Rodrigo Hernaiz Gomez and Marianne Mithun for further useful discussion and references. Above all, we thank all the language speakers who participated in this research for their friendship and help.

Abbreviations

1 = first person
2 = second person
3 = third person
A = agent
ACC = accusative
BEN = benefactive
COMP = complementiser
COP = copula
D = dependent
DEM = demonstrative
DESID = desiderative
DIS = disharmonic
DU = dual
EXCL = exclusive
FP = final particle
FUT = future
I = independent
IPFV = imperfective
IRR = irrealis
LOC = locative
NEG = negation
NOM = nominative
OBJ = object
PFV = perfective
PL = plural
POL = politeness marker
POSSD = possessed noun
PROG = progressive
PST = past
QUOT = quotative
R = realis
RR = reflexive/reciprocal
SBJ = subject
SEQ = sequential
SG = singular
TOP = topic

References

Aikhenvald, Alexandra Y. & R. M. W. Dixon. 2006. Introduction. In R. M. W. Dixon & Alexandra Y. Aikhenvald (eds.), *Complementation: a cross-linguistic typology*, 1–48. Oxford: Oxford University Press.

Barth, Danielle & Nicholas Evans (eds.). 2017a. The Social Cognition Parallax Corpus (SCOPIC). *Language Documentation and Conservation Special Publication* 12.

Barth, Danielle & Nicholas Evans. 2017b. The social cognition parallax corpus (SCOPIC): design and overview. In Danielle Barth & Nicholas Evans (eds.), *The Social Cognition Parallax Corpus (SCOPIC)* (Language Documentation and Conservation Special Publication 12). 1–21.

Barth, Danielle, Nicholas Evans, I Wayan Arka, Henrik Bergqvist, Diana Forker,

Sonja Gipper, Gabrielle Hodge, Eri Kashima, Yuki Kasuga, Carine Kawakami, Yukinori Kimoto, Dominique Knuchel, Norikazu Kogura, Keita Kurabe, John Mansfield, Heiko Narrog, Desak P. Eka Pratiwi, Saskia van Putten, Chikako Senge & Olena Tykhostup. 2021. Language vs. individuals in cross-linguistic corpus typology. In Stefan Schnell, Geoffrey Haig & Frank Seifart (eds.), *Doing corpus-based typology with spoken language corpora: State of the art* (Language Documentation & Conservation Special Publication 25). 1–56.

Barth, Danielle, Nicholas Evans, Sonja Gipper, Stefan Schnell, Henrik Bergqvist, Menguistu Amberber, I Wayan Arka, Christian Döhler, Diana Forker, Volker Gast, Dolgor Guntsetseg, Gabrielle Hodge, Eri Kashima, Yukinori Kimoto, Norikazu Kogura, Dominique Knuchel, Inge Kral, Keita Kurabe, John Mansfield, Heiko Narrog, Desak Putu Eka Pratiwi, Hiroki Nomoto, Seongha Rhee, Alan Rumsey, Lila San Roque, Andrea C. Schalley, Asako Shiohara, Elena Skribnik, Olena Tykhostup, Saskia van Putten & Yanti. 2024. The Social Cognition Parallax Interview Corpus (SCOPIC) Project Guidelines. In Danielle Barth & Nicholas Evans (eds.), *The Social Cognition Parallax Corpus (SCOPIC)* (Language Documentation and Conservation Special Publication 12). 163–237.

Brugman, Hennie & Albert Russel. 2004. Annotating multi-media/multi-modal resources with ELAN. In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa & Raquel Silva (eds.), *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, 2065–2068. Lisbon: European Language Resources Association (ELRA).

Charrad, Malika, Nadia Ghazzali, Véronique Boiteau & Azam Niknafs. 2014. NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, 61(6). 1–36.

Chomsky, Noam. 1957. *Syntactic structures.* The Hague: Mouton.

Deutscher, Guy. 2000. *Syntactic change in Akkadian.* Oxford: Oxford University Press.

De Villiers, Jill. 2000. Language and theory of mind: what are the developmental relationships?. In Simon Baron-Cohen, Helen Tager-Flusberg & Donald J. Cohen (eds.), *Understanding other minds: perspectives from developmental cognitive neuroscience* 2nd edn., 83–123. New York: Oxford University Press.

De Villiers, Jill G. & Peter A. De Villiers. 2003. Language for Thought: Coming to understand false beliefs. In Dedre Gentner & Susan Goldin-Meadow (eds.), *Language in Mind*. 335-384. Cambridge: MIT Press.

De Villiers, Jill G. & Peter A. De Villiers. 2014. The role of language in Theory of Mind Development. *Top Lang Disorders* 34(4). 313–328.

De Villiers, Jill G. & Jennie E. Pyers. 2002. Complements to cognition: the relationship between complex syntax and false-belief understanding. *Cognitive Development* 17(1). 1037–1060.

Diessel, Holger & Katja Hetterle. 2006. Causal clauses: a cross-linguistic investigation of their structure, meaning and use. In Peter Siemund (ed.), *Linguistic Universals and Language Variation*, 21–52. Berlin: Mouton de Gruyter.

Dingemanse, Mark, Andreas Liesenfeld, Marlou Rasenberg, Saul Albert, Felix K. Ameka, Abeba Birhane, Dimitris Bolis, Justine Cassell, Rebecca Clift, Elena Cuffari, Hanne De Jaegher, Catarina Dutilh Novaes, N. J. Enfield, Riccardo Fusaroli, Eleni Gregoromichelaki, Edwin Hutchins, Ivana Konvalinka, Damian Milton, Joanna Rączaszek-Leonardi, Vasudevi Reddy, Federico Rossano, David Schlangen, Joanna Seibt, Elizabeth Stokoe, Lucy Suchman, Cordula Vesper, Thalia Wheatley, Martina Wiltschko. 2023. Beyond Single-Mindedness: A Figure-Ground Reversal for the Cognitive Sciences. *Cognitive Science* 47(1). e13230.

Evans, Nicholas. 2006. Who said polysynthetic languages avoid subordination? Multiple subordination strategies in Dalabon. *Australian Journal of Linguistics* 26(1). 31–58.

Evans, Nicholas. 2021. Social cognition in Dalabon. In Danielle Barth & Nicholas Evans (eds.), *The Social Cognition Parallax Corpus (SCOPIC)* (Language Documentation and Conservation Special Publication 12). 22–84.

Evans, Nicholas, Francesca Merlan & Maggie Tukumba. 2004. *A first dictionary of Dalabon (Ngalkbon)*. Winnellie: Bawinanga Aboriginal Corporation.

Evans, Nicholas & David Wilkins. 2000. In the mind's ear: the semantic extensions of perception verbs in Australian languages. *Language* 76(3). 546–592.

Frajzyngier, Zygmunt. 1984. On the Origin of say and se as complementizers in Black-English and English-based creoles. *American Speech* 59(3). 207–210.

Givón, Thomas. 1991. The evolution of dependent clause morpho-syntax in Biblical Hebrew. In Elizabeth Closs Traugott & Bernd Heine (eds.), *Approaches to Grammaticalization: Volume II. Types of grammatical markers*, 257–310. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Grzech, Karolina. & Henrik Bergqvist, 2025. Epistemicity in language: current horizons, future directions. In Karolina Grzech & Henrik Bergqvist (eds.),

Expanding the Boundaries of Epistemicity: Epistemic Modality, Evidentiality, and Beyond, 1–30. Berlin: De Gruyter Mouton.

Harris, Alice & Lyle Campbell. 1995. *Historical syntax in cross-linguistic perspective*. Cambridge: Cambridge University Press.

Hernáiz Gomez, Rodrigo. 2024. The grammaticalization of manner expressions into complementizers: insights from Semitic languages. *Linguistics: An Interdisciplinary Journal of the Language Sciences* 62(3). 617–651.

Hodge, Gabrielle, Kazuki Sekine, Adam Schembri & Trevor Johnston. 2019. Comparing signers and speakers: Building a directly comparable corpus of Auslan and Australian English. *Corpora* 14(1). 63–76.

Kimoto, Yukinori, Asako Shiohara, Danielle Barth, Nicholas Evans, Norikazu Kogura, I Wayan Arka, Desak Putu Eka Pratiwi, Yuki Kasuga, Carine Kawakami, Keita Kurabe, Heiko Narrog, Hiroki Nomoto, Hitomi Ono, Alan Rumsey, Andrea C. Schalley, Yanti, Akiko Yokoyama. 2024. Syntactic embedding or parataxis? Corpus-based typology of complementation in language use. In Danielle Barth & Nicholas Evans (eds.), *The Social Cognition Parallax Corpus (SCOPIC)* (Language Documentation and Conservation Special Publication 12). 126–162.

King, Ronald S. 2015. *Cluster analysis and data mining: An introduction.* Dulles: Mercury Learning and Information.

Klamer, Marian. 2000. How report verbs become quote markers and complementisers. *Lingua* 110(2). 69–98.

Kolde, Raivo. 2019. pheatmap: Pretty heatmaps (R package version 1.0.12). Available online at: https://CRAN.R-project.org/package＝pheatmap (Accessed 2025.12.28).

Levshina, Natalia. 2022. Corpus-based typology: applications, challenges and some solutions. *Linguistic Typology* 26(1). 129–160.

Lucas, Carolina, Patrick Wong, Jon Klein, Tiago B. R. Castro, Julio Silva, Maria Sundaram, Mallory K. Ellingson, Tianyang Mao, Ji Eun Oh, Benjamin Israelow, Takehiro Takahashi, Maria Tokuyama, Peiwen Lu, Arvind Venkataraman, Annsea Park, Subhasis Mohanty, Haowei Wang, Anne L.Wyllie, Chantal B. F. Vogels, Rebecca Earnest, Sarah Lapidus, Isabel M. Ott, Adam J. Moore, M. Catherine Muenker, John B. Fournier, Melissa Campbell, Camila D. Odio, Arnau Casanovas-Massana, Yale IMPACT Team, Roy Herbst, Albert C. Shaw, Ruslan Medzhitov, Wade L. Schulz, Nathan D. Grubaugh, Charles Dela Cruz, Shelli Farhadian, Albert

I. Ko, Saad B. Omer & Akiko Iwasaki. 2020. Longitudinal analyses reveal immunological misfiring in severe COVID-19. *Nature* 584. 463–469.

Matsui, Tomoko, Hannes Rakoczy, Yui Mirua and Michael Tomasello. 2009. Understanding of speaker certainty and false-belief reasoning: a comparison of Japanese and German preschoolers. *Developmental Science* 12(4). 602–613.

Mayer, Thomas & Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.). *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*. 3148–3163. Reykjavik: European Language Resources Association (ELRA).

Mithun, Marianne. 2025. The Mighty Demonstrative. *Linguistic Typology at the Crossroads* 5-2. 104-122.

Noonan, Michael. 1985. Complementation. In Timothy Shopen (ed.), *Language typology and syntactic description, Vol. II, Complex Constructions*, 42–140. Cambridge: Cambridge University Press.

R Core Team. 2023. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ (Accessed 2025.12.28).

Reesink, Ger P. 1993. 'Inner speech' in Papuan languages. *Language and Linguistics in Melanesia* 24. 217–225.

Rumsey, Alan, John Mansfield & Nicholas Evans. 2022. The sound of one quotation mark: Quoted speech in Indigenous Australian narrative. In Alexandra Aikhenvald, Robert Bradshaw, Luca Ciucci & Pema Wangdi (eds.), *Celebrating Indigenous Voices. Legends and Narratives in Languages of the Tropics*, 33–72. Berlin: De Gruyter Mouton.

Saito, Hiroaki. 2021. Grammaticalization as decategorialization. *Journal of Historical Syntax* 5(10). 1–24.

San Roque, Lila, Alan Rumsey, Lauren Gawne, Stef Spronck, Darja Hoenigman, Alice Carroll, Julia Miller & Nicholas Evans. 2012. Getting the story straight: language fieldwork using a narrative problem-solving task. *Language Documentation and Conservation* 6. 134–173.

Sauerland, Uli, Bart Hollebrandse & František Kratochvil. 2020. When hypotaxis looks like parataxis: embedding and complementizer agreement in Teiwa. *Glossa:*

a journal of general linguistics 5(1). 89.

Schnell, Stefan, Geoffrey Haig & Frank Seifart. 2021. The role of language documentation in corpus-based typology. In Geoffrey Haig, Stefan Schnell & Frank Seifart (eds.), *Doing corpus-based typology with spoken language data: State of the art*, 1–28. Honolulu: University of Hawai'i Press.

Schnell, Stefan & Nils Norman Schiborr. 2022. Crosslinguistic Corpus Studies in Linguistic Typology. *Annual Review of Linguistics* 8(1). 171–191.

Spronck, Stef & Daniela Casartelli. 2021. In a manner of speaking: How reported speech may have shaped grammar. *Frontiers in Communication* 6. 624486.

Widmer, Manuel, Sandra Auderset, Johanna Nichols, Paul Widmer & Balthasar Bickel. 2017. NP recursion over time: Evidence from Indo-European. *Language* 93. 799–826.

Corpora and databases

A culturally informed corpus of Dalabon

Ponsonnet, Maïa. 2013. *A culturally informed corpus of Dalabon*. Endangered Language Archive. https://www.elararchive.org/dk0071/ (Accessed 2025.12.28).

Datenbank für Gesprochenes Deutsch

IDS, *Datenbank für Gesprochenes Deutsch (DGD)* [PF_E_00134_SE_01_T_01, FOLK_E_00337_SE_01_T_01, FOLK_E_00337_SE_01_T_02, FOLK_E_00337_SE_01_T_03, FOLK_E_00144_SE_01_T_01, PF_E_00016_SE_01_T_01, ZW_E_00979_SE_01_T_01]. http://dgd.ids-mannheim.de (Accessed 2025.12.28).

SCOPIC Corpus

Barth, Danielle & Nicholas Evans. 2024. *SCOPIC* 1.0 corpus files. SocCog-corp01 at catalog.paradisec.org.au. https://dx.doi.org/10.26278/1YH7-J821.

Sydney Speaks Corpus

Travis, Catherine E., James Grama, Simon Gonzalez, Benjamin Purser and Cale Johnstone. 2023. *Sydney Speaks Corpus*. ARC Centre of Excellence for the Dynamics of Language, Australian National University. https://dx.doi.org/10.25911/m03c-yz22.

The Yurakaré archive

van Gijn, Rik, Vincent Hirtzel, Sonja Gipper & Jeremías Ballivián Torrico. 2011. *The Yurakaré archive*. Online language documentation, DoBeS Archive, MPI Nijmegen. https://hdl.handle.net/1839/00-0000-0000-0016-662E-4.

CONTACT

nicholas.evans@anu.edu.au

Linguistic Typology at the Crossroads 5-2 (2025): 105-122 ISSN 2785-0943

The Mighty Demonstrative

MARIANNE MITHUN
UNIVERSITY OF CALIFORNIA, SANTA BARBARA

Submitted: 13/03/2024 Revised version: 27/02/2025
Accepted: 09/04/2025 Published: 25/02/2026

Abstract

Until recently, much typological work was necessarily based on comparisons of grammatical descriptions, still a fundamental resource. But as the field has progressed, it has become ever clearer that some aspects of structure emerge best in unscripted speech. In many languages, dependent clauses in complex sentence constructions are formed with demonstratives or their descendants. Examination of speech in Mohawk, an Iroquoian language of northeastern North America, might add detail to our understanding of a pathway of development that might not be obvious from elicited sentences or written materials. In this language, demonstratives are strikingly frequent in speech, often as placeholders while speakers formulate the next idea. Placeholders rarely appear in elicited examples in grammars or written materials, because speakers do not tend to think of them as integral parts of the language on a par with nouns and verbs. But the Mohawk placeholders are even more frequent than their counterparts in many other languages. Their pervasiveness has crystalized into a discourse construction in which speakers put forth one idea followed by a demonstrative in one intonation unit, then add elaboration in the next, often adding information about an argument of the first clause. This construction could be interpreted as equivalent to complement constructions in other languages, but a closer look indicates that it may represent a stage along one pathway of development.

Keywords: complement constructions; demonstratives; diachrony; discourse, placeholders; prosody

1. Introduction

Cross-linguistically, complex syntactic constructions are often based on demonstratives or related forms. Some examples of object complement constructions in English (eng; Indo-European, Germanic)[1] and German (deu; Indo-European, Germanic) are in (1) and (2).

(1) Object complements
  a. *I thought they said*
    *[**that** they put the case aside in Europe]*.
  b. *Ich dachte, sie hätten gesagt,*
    *[**dass** sie den Fall in Europa beiseite gelegt hätten]*.

(2) Object complements
  a. *They thought*
    *[**that** they would make us lose our language]*.
  b. *Sie dachten,*
    *[**dass** sie uns vielleicht dazu bringen würden, unsere Sprache zu verlieren]*.

Some examples of subject complement clauses are in (3) and (4).

(3) Subject complements
  a. ***It** made me happy*
    *[**that** he would learn this]*.
  b. *Es hat mich glücklich gemacht,*
    *[**dass** er das lernen würde]*.

(4) Subject complements
  a. ***It** is time*
    *[**that** I invite the public]*.
  b. ***Es** ist Zeit,*
    *[**dass** ich die Menschen einlade]*.

[1] Here and in what follows, when a language is named for the first time, the ISO 639-3 code and its genetic classification according to Glottolog are provided.

Indeed, Heine & Kuteva (2007: 225, 230) point out that cross-linguistically, demonstrative pronouns are common sources of complementizers.

2. Beyond Germanic

Apparently similar constructions also occur outside of Germanic languages. Languages of the Iroquoian family are indigenous to eastern North America. The Northern Iroquoian languages contain just three lexical categories defined in terms of morphological structure: verbs (V), nouns (N), and particles (P). They are polysynthetic: verbs in particular can show elaborate morphological structures, including noun incorporation. They are also head-marking: all verbs contain pronominal prefixes identifying their core arguments. Morphological verbs can accordingly serve as predicates, as in other languages, but also as complete clauses in themselves, as well as referring expressions. There is no case marking on nominals. Particles by definition have no internal structure, though they may be compounded. Examples here are from Mohawk (moh), a language of the Northern branch of the family, spoken in what is now Quebec, New York State, and Ontario. All are from unscripted connected speech, drawn from a corpus of approximately 70,000 words representing eighty speakers.

(5) Mohawk morphological and syntactic structure: Watshenní:ne' Sawyer, speaker

P	P	P	V
Sok	*nòn:wa'*	*tsi*	*iehiatonhseratahkwáhtha'*
sok	nonhwa'	tsi	ie-hiaton-hser-a-t-ahkw-aht-ha'
so	now	place	INDF.AGT-write-NMLZ-LK-be.in-INS-INS-HAB
So	then	place	one inserts letters with it

V
ionsaiakwakwátho'.
i-onsa-iakwa-kwatho-'
TRL-REP.FAC-1EXCL.PL.AGT-stop.by-PFV
we all went back there

'So then we went to the post office.'

There are three demonstratives, all of which can be used to refer on their own, or in apposition with a nominal. They do not distinguish number. The discourse demonstrative *né:* has also grammaticalized into a definite article *ne* 'the aforementioned'.

(6) Demonstratives

kí:ken, ki:	'this one, this, these'
thí:ken, thi:	'that one, that, those'
né:	discourse demonstrative (anaphoric or cataphoric)

Some examples of demonstratives are in (7), (8), and (9).

(7) Proximal: Joe Awenhráthon Deer, speaker

Ienskatahsónteren'	***kí:ken.***
i-en-s-k-at-ahsonteren-'	**kiken**
TRL-FUT-REP-1SG.AGT-MID-connect-PFV	**this**
'I'll continue **this**.'	

(8) Distal: Paul Deer, speaker

Tówa'	***thí:ken***	*tehontténions.*
towa'	***thi:ken***	te-hon-at-teni-ion-s
maybe	**that**	DV-M.PL.PAT-MID-change-DISTR-HAB

'Maybe **that** changes them.'

(9) Discourse anaphoric: Sonny Edwards, speaker

Né:	*iá:ken'*	*rón:kwe*	*ohniare'kó:wa*	*rotòn:'on.*
ne:	iaken'	r-onkwe	o-hniar-e' = kowa	ro-aton-'-on
that	HRS	M.SG-person	N-snake-NS = AUG	M.SG.PAT-became-INCH-ST

'**That one** was about the man who became a serpent.'

It would appear that these demonstratives mark complement constructions much like those in English and German. Some look like object complements.

(10) Object complement?

Í:kehre'	*ni'*	*rón:ton'*	*ken*
i-k-ehre-e'	n = i'i	ron-aton-'	ken
PROTH-1SG.AGT-think-ST	ART = 1	M.PL.PAT-say-ST	TAG

'I myself was thinking they said you know'

[thí:ken	***ne:***	*onhwentsakaiòn:ne*	*kénh*	*nekáːti*	*rotí:ien']*.
thiken	ne:	onhwentsi-akaion = hne	kenh	nekati	roti-ien-'
that	that	land-be.old = place	there	side	M.PL.PAT-set-ST

'[**that** they had put the case aside in Europe].'

(11) Object complement?

Ronnéhrahkwe'	*ónhte'*
ronn-hre-ahkwe'	onhte'
M.PL.AGT-think-PAST	perhaps

'They thought perhaps'

[thí:ken	*ahshakotiwennáhton'te'*
thiken	aa-hshakoti-wenn-ahton-'t-e'
that	OPT-3PL > 3PL-word-disappear-CAUS-PFV

ne	*onkwehonwehnéha']*.
ART	person = real = style
ne	onkwe = honwe = neha'

'[**that** they could make our language disappear].'

The examples in (12) and (13) appear to be subject complement constructions.

(12) Subject complement? Carol Phillips, speaker

Wakatshenón:ni
wak-atshen-onni
N > 1SG-happy-make

'**It** made me happy'

[kí:ken	*tsi*	*enhaweientéhta'ne'*	*wáhi'].*
kiken	tsi	en-ha-weiente-ht-a'n-e'	wahi'
this	how	FUT-M.SG.AGT-know.how-CAUS-INCH-PFV	TAG

'[**that** he would learn this you know].'

(13) Subject complement? Paul Deer, speaker

Ó:nen	*á:re'*	*ki:*	*sakáhewe'*
onen	are'	kiken	sa-ka-hew-e'
now	again	this	REP-N.AGT-carry-ST

'**It**'s time again'

[kí:ken	*wà:kehre'*	*ensakhehón:karon'*
kiken	wa'-k-hre-e'	en-sa-khe-honkaron-'
this	FAC-1SG.AGT-think-PFV	FUT-REP-1SG > 3PL-invite-PFV

ne	*onkwe'tá:ke].*
ne	onkwe't-ake
ART	person-be.multiple

'[**that** we invite the public I think].'

So do we have a universal? Things become more interesting if we look closely at what speakers actually say, complete with prosody and context.

3. Placeholders

Hayashi & Yoon (2010) and Podlesskaya (2010) describe placeholders, lexical items that fill the slot of a delayed word or constituent as speakers search for words. They note that most placeholders develop out of demonstrative pronouns. Because speakers usually do not consider them an integral part of language, they were difficult to capture before easy access to audio and video recording technology.

Mohawk demonstratives occur pervasively in this function. From here on, examples are arranged such that each intonation unit or prosodic phrase begins on a separate line. As characterized by Chafe (1987: 22), "An intonation unit is a sequence of words combined under a single, coherent intonation contour, usually preceded by a pause". When an intonation unit is longer than a line, it is indented on the following line.

Punctuation at the ends of lines reflects prosody: a dash signals a truncation, a comma the end of a non-final intonation unit, and a period a terminal fall in pitch at the end of a sentence-final intonation unit. In (14) the speaker was searching for a word to characterize the translation work they were doing.

(14) Placeholder for word search: Charlotte Kaherákwas Bush, speaker

Tóka'	*né:*	*ki'*	*kí:ken*	*ionkwaiió'te'*	*wáhi,*
toka'	ne:	ki'	kiken	ionkwa-io't-e'	wahi'
maybe	that	in.fact	this	1PL.AGT-work-ST	TAG

'Maybe the work we're doing'

kí:ken--
'this'

nahò:ten'	*kati'*	*kí:*	*tetewawennatá—*
n-a-h-o'ten-'	kati'	kiken	te-tewa-wenn-a-ta—
PRT-FAC-N-be.a.kind.of-PFV	actually	this	DV-1INCL.PL.AGT-word-LK-X

'is it actually--'

tetewawennanetáhkwahs	*ken*	*káton,*
te-tewa-wenn-a-net-ahkw-ahs	ken	katon
DV-1INCL.PL.AGT-word-LK-layer-REV-HAB	Q	or

'unlayering the words or,'

tetewawennaténie's?
te-tewa-wenn-ateni-e's
DV-1INCL.PL.AGT-word-change-HAB
'changing the words?'

In (15) the speaker was searching for a way to refer to drugs.

(15) Placeholder for word search: Joe Awenhráthon Deer, speaker
['Are you talking about liquor?']

Én:. Tetsá:ron ki:,
'Yes. Both these,'

ohné:ka' tánon' ne'ne--
o-hnek-a' tanon' ne'ne
N-liquid-NS and that which
'liquor and that which—'

thi:kén:--
'**that**'

rati'niónhsakon *ronneta's* . . .
rati-'nionhs-akon ronn-et-a'-s
M.PL.INAL.POSS-nose-interior M.PL.AGT-be.in-CAUS-HAB
'stuff they put in their nose.'

In (16) the speaker was searching for a term for 'streetcar'

(16) Placeholder for word search: Doris White, speaker
['We'd cross the bridge']

ok *tóka'* *takatákhe'* ***ne'--***
ok toka' ta-ka-takh-e' ne'e
and.then maybe CSL.FAC-N.AGT-run-ST that.is
'and then maybe it would come running toward us—'

a:, né: ki' *ne* *wáhi'* ***ne:*** *a:, a:,*
ah that in.fact ART TAG **that** ah ah

streetcar *iakwana'tónhkwahkwe'* *wahón:nise'.*
iakwa-na'ton-hkw-ahkwe' wa-h-onnis-e'
1EXCL.PL.AGT-call-INS-PAST.HAB FAC-N-be.long-PFV

'a streetcar we used to call it long ago.'

4. Prosodic structure

As in other languages, speakers tend to speak in spurts, introducing one new idea at a time per intonation unit (Chafe 1987, 1994 and elsewhere). This prosodic pattern can be seen in (17), as the speaker announced that 'A woman named Kahentoréhtha got worse and was taken to Chateauguay'. A number of these intonation units end with a demonstrative *kí:ken* 'this' or its shorter form *ki:*, or a simple filler like *a:* 'ah' or *en:* 'um'.

(17) Intonation units: Joe Tiohrakwén:te' Dove, speaker

Tseià:ta	***ki:kén:***--
ts-ie-a't-at	**kiken**
REP-FI.AGT-body-be.one	**this**

a:,
ah,

taiakorihwà:rekshe'	*ó:nen*	***ki:,***
ta-iako-rihw-a-hreksh-e'	onen	**kiken**
CSL.FAC-FI.SG.PAT-matter-LK-push-PFV	now	**this**

Shahrè:'on	*iahshakotí'teron',*
Shahre'on	i-a-hshakoti-i'teron-'
PLACENAME	TRL-FAC-3PL > FI-reside-PFV

ó:nen	***ki:***,
onen	**kiken**
now	**this**

né:ne	*Kahentoréhtha'*	*konwá:iats*	*ki:.*
nene	ka-hent-oreht-ha'	Ia-iat-s	kiken
that.one	FZ.AGT-field-balance-HAB	FI > FZ.SG-call-HAB	this

'A woman named Kahentoréhtha' got worse and was taken to Chateauguay.'

The woman, a brand new topic, was introduced in the first intonation unit, followed by the placeholder *kí:ken* 'this' then a pause, before the next intonation unit which consisted of just the filler 'ah' as the speaker formulated his comment. That new piece of information, 'she was taken to Chateauguay', was uttered in an intonation unit of its own. The name of the woman was given in a final intonation unit. (The final demonstrative is part of a different construction. It is not a placeholder, but rather an antitopic, recapitulating the discourse topic).

Similar prosodic structure can be seen in (18), 'Another time, chickens were brought in to breed.' The speaker first introduced a shift in time followed by a filler 'ah', then in the next intonation unit the event 'they brought them', then another filler, and finally the referent 'chickens to breed'.

(18) Intonation units: Joe Awenhráthon Deer, speaker

Ó:ia's	*ni'*	***a:,***
oia'=s	ohni'	**a:**
other=DISTR	also	**ah**

wahatíhewe',
wa-hati-hew-e'
FAC-M.PL.AGT-bring-PFV

en:,
um

kítkit	*thaontenahskón:ní'*	*wáhe'.*
kitkit	th-a-w-ate-nahskw-onni-'	wahe'
chicken	CONTR-FAC-N.AGT-MID-domestic.animal-make-PFV	TAG

'Another time chickens were brought in to breed.'

This speaker followed that introduction with the comment in (19), 'Maybe you remember, the building is still there behind Kawinéhtha's place.'

(19) Intonation units: Joe Awenhráthon Deer, speaker

Ta' nòn:wa' *sè:iahre',*
towa' = nonhwa' s-ehiar-e'
maybe = now 2SG.AGT-remember-ST

shé:kon *tkanónhsote'* ***thi:kén:,***
shekon t-ka-nonhs-ot-e' **thiken**
still CSL-N.AGT-house-stand-ST **that**
énska *ohnà:ken'* ***thi:kén:,***
one behind **that**

a:, *né:* *ki:* ***né:,***
ah that.is this **that**

Ka'-- *en:,* *Kahon'--* *Kawinehthà:ke* *ohna:kén:,*
ka'-- ah Kahon'-- Kawinéhtha's place behind

énska *shé:* *kanónhsote'.*
enskat shekon ka-nonhs-ot-e'
one still N.AGT-house-stand-ST
'Maybe you remember, the building is still there behind Kawinéhtha's place.'

Again each new idea was set out in a separate intonation unit, often preceded by a filler or demonstrative placeholder and pause, as the speaker formulated what he was going to say.

This speaker then continued with (20), 'That's where the chickens live'.

(20) Intonation units: Joe Awenhráthon Deer, speaker

Thos *non:* *konti'teróntonhkwe'* ***ki:kén:,***
tho shes nonwe' konti-'teront-on-hkwe' **kiken**
there CUST place FZ.PL.AGT-reside-ST-PAST **this**

kítkit.
chicken
'That's where the chickens live.'

It is important to note that these are all highly skilled first-language speakers; the pausing does not in any way reflect a lack of facility with the language.

The demonstratives at the ends of intonation units (apart from the sentence-final antitopics) have distinctive prosody. Basic demonstratives have penultimate stress: *kí:ken* 'this, these', *thí:ken* 'that, those'. But in these contexts, the pitch continues to rise into the final syllable, which is lengthened: *ki:kén:, thi:kén:*. The difference can be seen in the pitch traces in Figures 1 and 2, a comparison of the demonstrative *kí:ken* in its basic use from (7), and *ki:kén:* in its placeholder use from (17).

(7) Basic demonstrative
Ienskatahsónteren' kí:ken.
'I should continue this.'

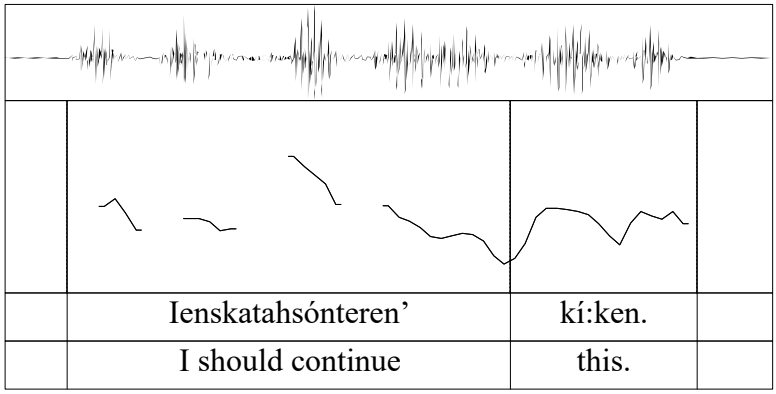


Figure 1: Basic demonstrative

(17) Placeholder demonstrative
Tseià:ta ***ki:kén:,***
one woman **this**

a:, taiakorihwà:rekshe' ó:nen ki:,
ah, she got worse now this,

Shahrè:'on iahshakotí'teron', . . .
Chateauguay there they placed her
'One woman, she got worse and was taken to Chateauguay.'

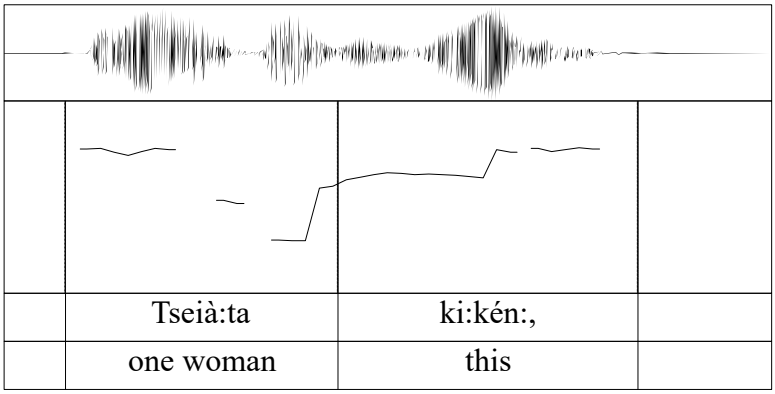


Figure 2: Placeholder demonstrative

5. Conventionalized discourse structure

The pattern of an intonation unit conveying one new idea ending in a demonstrative with rising pitch signaling that further detail is to follow has crystallized into a conventional discourse structure. The further detail is not limited to a specific syntactic role; it may elaborate on a participant, place, time, reason, and more. In (21), the speaker first elaborated on what was numerous, namely 'stuff causing us worry', then the location.

(21) Statement + elaboration: Joe Aweenhráthon Deer, speaker

A:	*nia'té:kon*	***ki:,***
a:	n-i-a'-te-ka-I	**kiken**
ah	PRT-TRL-FAC-DV-N.AGT-be.amount	**this**
ah	it is many	**this**

tsi,	*nia'té:kon*	*tiótkon*	*teka'nikónhrhare'*	***ki:kén:,***
tsi	n-i-a'-te-ka-i	t-io-kon	te-ka-'nikonhr-har-e'	**kiken**
ah	PRT-TRL-FAC-DV-N.AGT-be.amount	always	DV-N.AGT-mind-hang-ST	**this**

nonkwaná:takon.
ne = onkwa-nat-akon
ART-1AL.POSS-town-interior
'There's a lot of stuff is causing us worry in our town.'

6. Standard complex syntax?

If we look back at the prosody of examples seen earlier translated with standard complement constructions, we can see that they show the same intonation patterning as the discourse structure of a statement ending in a demonstrative followed by elaboration. Here the same examples are displayed with their prosody, each line representing an intonation unit.

(10) Object complement? Skawén:nati Montour, speaker

Í:kehre'	*ni'*	*rón:ton'*	*ken*	***thí:ken,***
i-k-hre-e'	n = i'i	ron-aton-'	ken	thiken
PROTH-1SG.AGT-think-ST	ART = 1	M.PL.PAT-say-ST	TAG	**that**

ne:,
that

onhwentsakaiòn:ne	*kénh*	*neká:ti*	*rotí:ien'.*
onhwentsi-akaion = hne	kenh	nekati	roti-ien-'
land-be.old = place	there	side	M.PL.PAT-set-ST

'I myself was thinking they said they had put the case aside in Europe.'

The pitch trace in Figure 3 shows that the demonstrative *ne:* was pronounced with rising pitch and increased length, then followed by a pause.

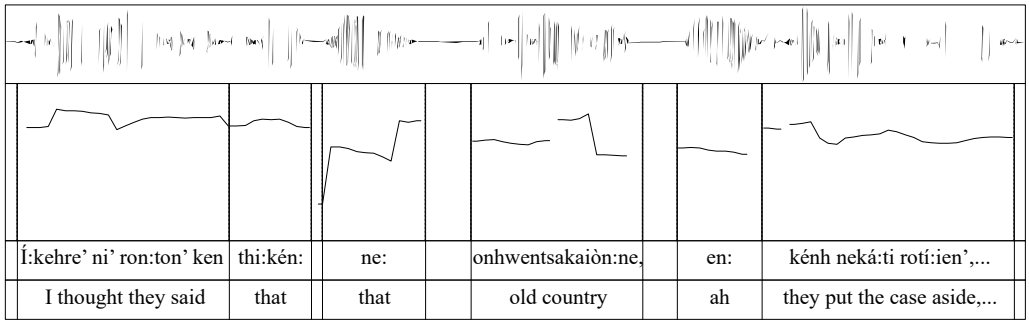


Figure 3: Prosodic structure of (10)

The same pattern can be seen in the other examples translated with complement constructions.

(11) Object complement? Dennis Stock, speaker

Ronnéhrahkwe'	*ónhte'*	***thí:ken,***
ronn-her-ahkwe'	onhte'	**thiken**
M.PL.AGT-think-PAST	perhaps	**that**

'They thought perhaps'

ahshakotiwennáhton'te'	*ne*	*onkwehonwehnéha'.*
aa-hshakoti-wenn-ahton-'t-e'	ne	onkwe=honwe=neha'
OPT-3PL>3PL-word-disappear-CAUS-PFV	ART	person=real=style

'they could make our language disappear.'

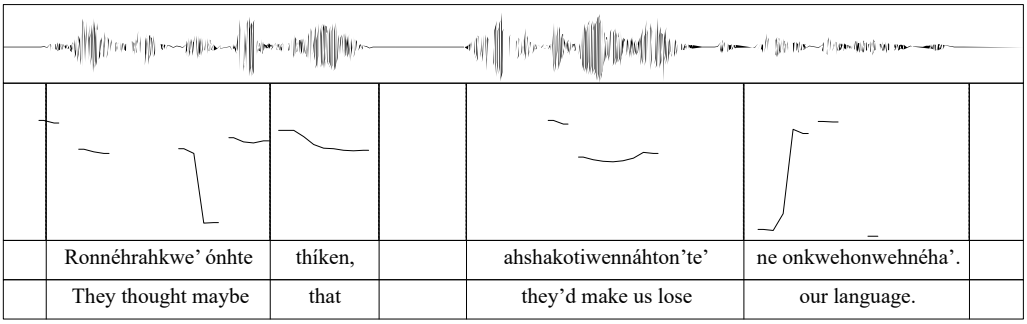


Figure 4: Prosodic structure of (11)

(12) Subject complement? Carol Phillips, speaker

Wakatshenón:ni	***kí:ken***	*tsi,*
wak-atshen-onni	**kiken**	tsi
N>1SG-happy-make	**this**	how

'**It** made me happy'

enhaweientéhta'ne'	*ki'*	*wáhi'*.
en-ha-weiente-ht-a'n-e'	ki'	wahi'
FUT-M.SG.AGT-know.how-CAUS-INCH-PFV	in.fact	TAG

'that he would learn this you know.'

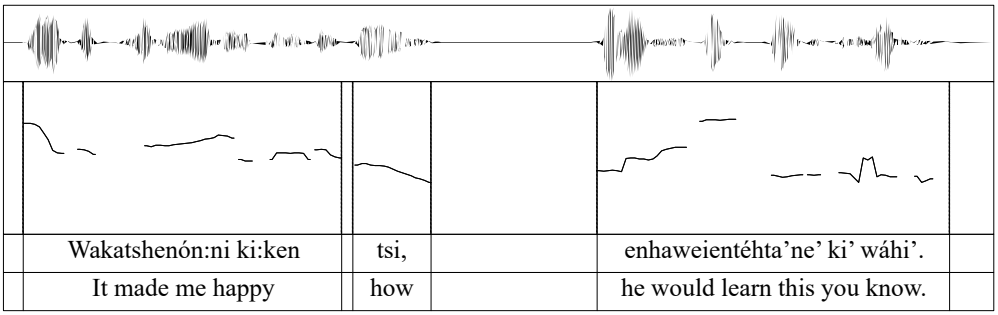


Figure 5: Prosodic structure of (12)

(13) Subject complement? Paul Deer, speaker

Ó:nen	*á:re'*	*ki:*	*sakáhewe'*	***kí:ken,***
onen	are'	kiken	sa-ka-hew-e'	**kiken**
now	again	this	REP-N.AGT-carry-ST	this

wà:kehre'	*ensakhehón:karon'*	*ne*	*onkwe'tà:ke.*
wa'-k-hre-e'	en-sa-khe-honkaron-'	ne	onkwe't=ake
FAC-1SG.AGT-think-PFV	FUT-REP-1SG>3PL-invite-PFV	ART	person=place

'**It**'s time again **that** we invite the public I think.'

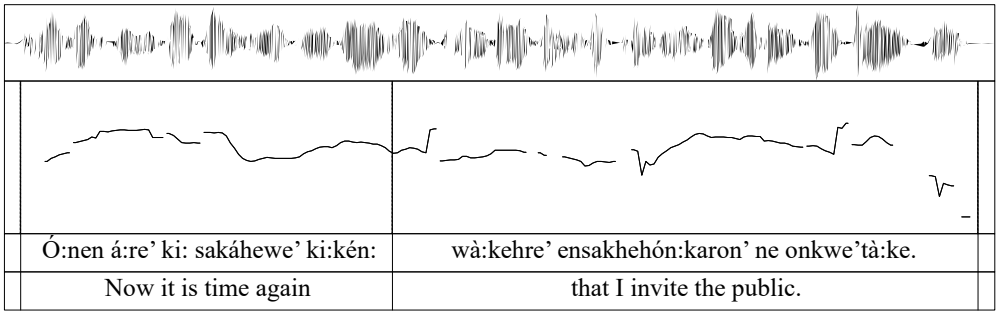


Figure 6: Prosodic structure of (13)

Interestingly, in some examples, the prosody of the intonation-unit final demonstrative is somewhat attenuated. In (11), 'They thought that they'd make us lose our language', and (13) 'Not it is time again that I invite the public', the final syllable of the demonstrative does not show as steep a rise in pitch as it does in some other instances of the construction. In (13), there was also little pause between the demonstrative and the following clause. Such patterns of course vary by speaker. They

could, however, be indicative of incipient grammaticalization of a syntactic complement construction.

7. Conclusion

Mohawk contains numerous sentences that are translated with English complement constructions. A broader look at language in use, however, shows that they are instances of a larger discourse construction which shapes the flow of information. These have apparently evolved out of a common pattern cross-linguistically whereby demonstratives are used as placeholders while the speaker searches for a word or the best way to formulate the next idea. This use can be seen in Mohawk, but the pattern has moved beyond: in perhaps the majority of cases, speakers know well what is to follow. An intonation unit-final demonstrative with continuing prosody indicates that more is to come, and of course serves as a floor-holding device.

Has this pattern evolved completely into complement constructions? It is not limited to elaboration of a core argument of a main clause: it can provide further details about time, place, reason, and more. There are some instances, however, where the final high pitch and length on the demonstrative, and/or the prosodic break between the demonstrative and the following material is somewhat attenuated. This could suggest that such discourse patterns are a stage along one pathway to the grammaticalization of complex syntactic structures.

Ultimately, typology can be richer and more interesting if we are not restricted to elicited translations of single sentences, out of context, from a contact language. Here the discourse context and prosody of unscripted speech reveal that what might at first appear to be canonical complex syntactic constructions are actually cases of a more general discourse construction. At the same time, they could add detail to our understanding of one pathway by which such constructions might develop.

Abbreviations

1 = 1st person
2 = 2nd person
3 = 3rd person
AGT = grammatical agent
ART = article
FUT = future
FZ = feminine-zoic
HAB = habitual aspect
HRS = hearsay
INAL = inalienable
OPT = optative
PAT = grammatical patient
PFV = perfective
PL = plural
POSS = possessive

AUG = augmentative
CAUS = causative
CSL = cislocative
CONTR = contrastive
CUST = customary
DISTR = distributive
DV = duplicative
EXCL = exclusive
FAC = factual
FI = feminine-indefinite
INCH = inchoative
INCL = inclusive
INDF = indefinite gender
INS = instrumental applicative
LK = linker
M = mansculine
MID = middle
N = neuter
NMLZ = nominalizer
NS = noun suffix
PRT = partitive
REP = repetitive
REV = reversive
SG = singular
ST = stative aspect
TAG = tag question
TRL = translocative
X = unanalyzable

Orthographic conventions

The symbol <i> is a glide [j] before vowels, the digraphs <en> and <on> nasalized vowels [ą] and [ų], and the apostrophe <’> glottal stop [ʔ]. The colon <:> indicates vowel length. The acute accent <ˊ> represents stress with high or rising tone, and the grave accent <ˋ> stress with falling tone.

References

Chafe, Wallace. 1987. Cognitive constraints on information flow. In Russell Tomlin (ed.), *Coherence and Grounding in Discourse*, 21–52. Amsterdam: John Benjamins.

Chafe, Wallace. 1994. *Discourse, Consciousness, and Time.* Chicago: University of Chicago Press.

Hayashi, Makoto & Kyung-Eun Yoon. 2010. A Cross-Linguistic Exploration of Demonstratives in Interaction. With Particular Reference to the Context of Word-Formulation Trouble. In Nino Amiridze, Boyd H. Davis, & Margaret Maclagan (eds.), *Fillers, Pauses and Placeholders*, 33–66. Amsterdam: John Benjamins.

Heine, Bernd & Tania Kuteva. 2007. *The Genesis of Grammar: A Reconstruction*. Oxford: Oxford University Press.

Podlesskaya, Vera. 2010. Parameters for Typological Variation of Placeholders. In Nino Amiridze, Boyd H. Davis, & Margaret Maclagan (eds.), *Fillers, Pauses and Placeholders*, 11–32. Amsterdam: John Benjamins.

CONTACT
mithun@linguistics.ucsb.edu

Linguistic Typology at the Crossroads 5-2 (2025): 123-160 ISSN 2785-0943

That's what I need: A multimodal study of Hebrew 'Reversed Pseudo-Clefts'

YAEL MASCHLER, HILLA POLAK-YITZHAKI
UNIVERSITY OF HAIFA - ISRAEL

Submitted: 12/02/2025 Revised version: 07/08/2025
Accepted: 10/12/2025 Published: 25/02/2026

Abstract

Employing Interactional Linguistic methodology and multimodal interactional analysis, we investigate the Hebrew [*ze ma she-* 'this is what' + clause] structure, known as a 'reversed pseudo-cleft'. The corpus consists of 9 hours of video-recorded casual conversation among friends and relatives, manifesting 70 [*ze ma she-* + clause] tokens. Instead of the traditional grammatical analysis of this structure as consisting of a nominalized clause functioning as predicate, embedded within a matrix clause, we argue for an analysis of *ze ma she-* as a 'fixed chunk', a construction which has grammaticized from repetitive discourse actions to serve particular functions in interaction. Furthermore, we argue that there is no justification for viewing this structure as a 'reversed' form of a Hebrew pseudo-cleft. We support our claims with evidence from prosodic, lexico-semantic, syntactic, pragmatic, and embodied patterning of the [*ze ma she-* + clause] tokens found in our corpus. We show that the structure functions in 71% of the instances for (1) framing prior talk metalingually or (2) claim-backing. Other, less frequent uses include (3) seeking clarification, (4) postulating some general truth, (5) disclaiming responsibility, and (6) getting back to a previous topic. Only two tokens throughout our data display (7) the summative function, claimed as by far the most common function for English reversed pseudo-clefts. Our study supports a view of grammar as a temporally-unfolding, tightly interwoven with embodied conduct, ever-evolving resource for carrying out social actions in the dialogical process of interaction.

Keywords: reversed pseudo-clefts; interactional linguistics; multimodality; Hebrew syntax; embodied syntax; fixed fragments.

1. Introduction

In the context of the special issue "Naturally Occurring Data in and beyond Linguistic Typology" of which this article is a part, we would like to argue that in the study of language, it is indeed necessary – but not sufficient – to focus on naturally occurring data. We must approach this type of data with a specific methodology and from a particular theoretical framework. Interactionally oriented approaches to language such as "Emergent Grammar" (Hopper 1987), "on-line syntax" (Auer 2009), and "dialogical grammar" (Linell 2009; Du Bois 2014) have brought about a radical change in our conceptualization of language from a hierarchical, autonomously-structured mental construct (Saussure 1959[1913]) to a usage-based, temporally-unfolding, tightly-interwoven with embodied conduct, ever-evolving resource for carrying out social actions in the dialogical process of interaction. Embracing these interactional approaches, we focus here on Hebrew[1] structures of the type:[2]

(1) ***ze*** ***ma*** ***she***-*'ani* *tsarix.*
this.M.SG what that-I need.PRS.M.SG
'that's what I need.'

(2) ***ze*** ***ma*** ***she***-*'ani* *rotse* *lehaspik,*
this.M.SG what that-I want.PRS.M.SG get_done.INF
'that's what I want to accomplish,'

(3) ***ze*** ***ma*** ***she***-*limdu* *'et-xem?*
this.M.SG what that-teach.PST.3M.PL ACC-3M.PL
'that's what they taught you?'

This structure, known as a 'reversed pseudo-cleft' – in Hebrew, *mishpat mevuka mehupax* (Azar 1992: 94) –, is composed of the masculine singular demonstrative pronoun *ze* 'this', followed by the interrogative question word *ma* 'what' and the complementizer (and general Hebrew subordinator) *she-* 'that' – all followed by a clause:

[1] Heb; Afro-Asiatic; Semitic.

[2] Unless otherwise noted, all Hebrew examples come from our data, see below. For transcription conventions, see Appendix.

(4) ***ze*** ***ma*** ***she-*** *[clause]*
this.M.SG what that-
'That's what [clause]'

According to traditional Hebrew grammar, such structures belong in the realm of complex syntax and can be analyzed as a nominal (or copular) matrix clause in which *ze* 'this' would be considered subject and the following '*ma she-* + [clause]' complex would be considered a nominalized clause functioning as predicate, embedded in the matrix clause. However, we would like to suggest viewing *ze ma she-* as **a fixed chunk** (Bybee 2003: 603), a construction which has grammaticized from repetitive discourse actions to serve particular functions in interaction. Furthermore, we will show that there is no justification in the data for perceiving this structure as a 'reversed' form of a pseudo-cleft.

Following a section on data and methodology (Section 2), we provide a survey of previous studies of related structures in several languages (Section 3). In Section 4 we present the lexico-semantic, syntactic, and prosodic patterning associated with the [*ze ma she-* + clause] structure found in our data. In Section 5 we turn to the heart of this study, a multimodal interactional analysis of the structure, investigating its functions in Hebrew casual conversation. Section 6 summarizes our findings and discusses their implications.

2. Data and methodology

Our data come from the Haifa Multimodal Corpus of Spoken Hebrew (Maschler et al. 2024). At the time of data collection, the corpus consisted of 9 hours of video-recorded casual conversation among friends and relatives, 2–6 participants per interaction, altogether 16 informal interactions between 42 different interlocutors, recorded during the years 2016–2019.

We manually searched for the *ze ma she-* construction throughout the data and found 70 tokens. These 70 tokens constitute the database for the present study.

The 'equivalent' English structure, sometimes termed a 'reverse(d) pseudo-cleft' (see Section 3 below) allows for a variety of question words to follow the '*that's*' part of the 'equivalent' structure, as in:

(5) *That's where Cathy Reid is.* (Küttner 2020: 252)

(6) *That's how Linda got her job in Puerto Rico.* (Küttner 2020: 262)

However, the only question word found in the Hebrew structure throughout our data is *ma* 'what'.

We included tokens with an adverb or a quantifier preceding the question word, such as:

(7) ***ze*** ***bediyuk*** ***ma*** ***she***-*'ani* *'omeret.*
this.M.SG exactly what that-I say.PRS.F.SG
'that's **exactly** what I'm saying.'

(8) ***ze*** ***kol*** ***ma*** ***she***-*mikro* *yode'a* *la'asot.*
this.M.SG all what that-microwave know.PRS.M.SG do.INF
'that's **all** a microwave can do.'

Altogether, 9 tokens of this kind were found among the 70 cases in our corpus.

Fully embracing the temporality and contingency of the moment-by-moment incremental unfolding of interaction (Hopper 1987; 2011; 2020), we employ the methodology of Interactional Linguistics – the study of language and languages in social interaction (Couper-Kuhlen & Selting 2018) – to study the ways in which the grammatical resource of the [*ze ma she-* + clause] structure is employed to produce social actions. To analyze the participants' embodied conduct involved in production of this target structure, we employ multimodal interaction analysis (Goodwin 2000, 2018; Mondada 2006).

3. Previous research on reversed pseudo-clefts and related structures

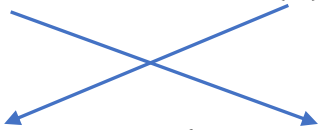
In the literature about 'equivalent' structures in European languages (mostly English, but there is some research also on German, Italian, and Norwegian), the structure is known as a 'reverse(d) pseudo-cleft' or a 'reverse *wh*-cleft' (e.g., Geluykens 1984; Erdmann 1986; Collins 1991; Oberlander & Delin 1996; Weinert & Miller 1996; Lambrecht 2001; Traugott 2008; De Cesare 2014) because of its ties with the pseudo-cleft structure.[3] We restrict our study to what has been characterized as 'headless'

[3] Some linguists distinguish between reversed pseudo-clefts and demonstrative reversed pseudo-clefts by terming the latter construction 'the DEM-BE-WH construction' (Ball 1991), 'demonstrative *wh*-clefts' (Biber et al. 1999: 961), 'demonstrative clefts' (Calude 2008), 'Type 2 reverse pseudo-cleft sentences'

reversed pseudo-clefts, i.e., those opening with a demonstrative – the most common type of reversed pseudo-cleft found in studies of the 'equivalent' English structure (e.g., Erdmann 1986; Collins 1991; Weinert & Miller 1996; Oberlander & Delin 1996; Calude 2008).[4]

An English pseudo-cleft, according to traditional grammar, constitutes "a Subject Verb Complement sentence with a *wh*-nominal clause as subject or complement" (Quirk et al. 1985: 1387–1389), as in (9a):

(9)	**A**	**COP**	**B**	
	Subject	**Verb**	**Complement**	
(9a)	*What he's done*	*is*	*(to) spoil the whole thing.*	(Quirk et al. ibid.)



	Complement	**Verb**	**Subject**
(9b)	*That*	*is*	*what he's done.*

The parallel reversed pseudo-cleft (9b) would consist of the same components – subject, verb, complement – in a reversed order, only with 'that' replacing the complement ('to spoil the whole thing'). In both cases – the pseudo-cleft and the reversed pseudo-cleft – we have an 'A copula B' structure.

A Hebrew pseudo-cleft structure (Azar 1992; Yatsiv-Malibert 2009; Maschler & Fishman 2020; Maschler & Pekarek Doehler 2022) consists of the *ma she-* 'what that' construction in part A, the demonstrative *ze* (or *hu*, the masculine singular personal pronoun) as copula, and a *that*-clause as part B, as in (10a):

(10)	**A**	**COP**	**B**
(10a)	***ma she-****kara*	*ze*	*she-[clause]*
	what that happen.PST.3M.SG	this.M.SG	that [clause]
	'what happened is that [clause]'		

(Garassino 2014), the '*that's what* construction' (Johansson 2001), or 'the [*that's wh*-clause] format' (Küttner 2020).

[4] Thus, we will not be dealing with cases in which part A carries a fuller lexical form, such as *the thing that…, the reason that…, the one that…* (cf., Ball 1991; Oberlander & Delin 1996: 197) or a NP as in examples such as ***Champagne*** *is what I like* (cf. De Cesare 2014: 10).

And the so-called 'reversed pseudo-cleft' structure consists of the demonstrative pronoun *ze* 'this'[5] in part A (instead of part B of the *pseudo-cleft*), no copula (Hebrew allows for non-copular nominal clauses), and the *ma she-* construction in part B, as in (10b):

	A	**COP**	**B**	
(10b)	*ze*	*Ø*	***ma***	***she**-kara*
	this.M.SG		what	that happen.PST.3M.SG
	'that's what happened.'			

According to Azar, such a reversed pseudo-cleft is a reduced form of a fuller sentence, in which the demonstrative *ze* 'this' stands for a full clause or infinitival phrase. The example he gives is:

(11a)	*hem*	*rotsim*	*le'exol*	*'otanu*	*bli*	*melax.*
	they	want.PRS.M.PL	eat.INF	us	without	salt
	'they want to eat us alive.'					

ze	*ma*	*she-hem*	*rotsim.*
this.M.SG	what	that they	want.PRS.M.PL

'that's what they want.' (adapted from Azar (1992: 94), our translation)

According to him, this reversed pseudo-cleft stands for the fuller form:

(11b)	**A**				**COP**	***B***		
	le'exol	***'otanu***	***bli***	***melax***	*ze*	*ma*	*she-hem*	*rotsim.*
	eat.INF	us	without	salt	this.M.SG	what	that they	want.PRS.M.SG
	'to eat us alive is what they want.'							

In traditional grammar and Information Structure approaches, the function of the pseudo-cleft is considered to be the backgrounding of material in part A in order to focus on part B (e.g., Jespersen 1949; Prince 1978; Geluykens 1988; Collins 1991; Azar 1992; Biber et al. 1999; Lambrecht 2001; Yatsiv-Malibert 2009, see literature review in Maschler & Pekarek Doehler 2022). However, it's been shown, in Interactional Linguistic studies on several languages, including Hebrew, that

[5] There is no *this* vs. *that* distinction in the Hebrew demonstrative system, thus *ze* can be translated as either one. When *ze* appears in a Hebrew reversed pseudo-cleft, we translate it with 'that', the more common demonstrative in English reversed pseudo-clefts (Weinert & Miller 1996: 181; Oberlander & Delin 1996: 189; Calude 2008: 79).

focusing, if it occurs, is a by-product of a different function, an epiphenomenon (Hopper 2001: 111). A more rigorous account of these structures is that part A functions as a *projecting construction* (Auer 2005; Günthner 2011), delaying the production of part B for various cognitive and interactional reasons. It also functions to frame (Goffman 1981) the discourse to come in particular ways (e.g., Hopper 2001; 2004; Hopper & Thompson 2008; Günthner & Hopper 2010; Günthner 2011; Pekarek Doehler 2011; Maschler & Fishman 2020; Maschler & Pekarek Doehler 2022).

In a special issue of *Lingua*, Maschler et al. (2023) compared the use of pseudo-cleft constructions in talk-in-interaction across six languages. We found that even though the languages are often quite different typologically from each other, and for some languages there is very little documentation of pseudo-clefts (Japanese, Mandarin, Swedish, Estonian), there are many similarities in the ways such structures are employed in interaction in those languages, suggesting possible universal interactional motivations for the grammatical properties of pseudo-clefts across languages.

As for the functions of reversed pseudo-clefts, they, too, have been studied within Information Structure approaches. According to most studies, the demonstrative pronoun denotes Given information, while the clause, which contains the prosodic nucleus of the tone unit, denotes New information (e.g., Stubbs 1983; Collins 1991; Weinert & Miller 1996[6]; Oberlander & Delin 1996[7]; Lambrecht 2001; De Cesare 2014). Oberlander & Delin (1996) found that the demonstrative is in many cases anaphoric but that it can also be cataphoric (cf. Collins 1991; Erdmann 1986; Calude 2008) or exophoric (cf. Calude 2008), and that the demonstrative may refer to long discourse segments (cf. Garassino 2014; Johansson 2001; Calude 2008). Similarly, Weinert and Miller (1996: 173) found that the demonstrative can refer to an entity in the immediately surrounding discourse or situation, allowing the speaker to hold onto previous ideas before moving on, and thus "creating a slight backwards pull" (cf. Erdmann 1986: 854).

[6] However, Weinert & Miller found that a large majority of reversed pseudo-clefts in their data manifest a stressed element in both part A and part B (1996: 189).

[7] However, Oberlander & Delin also found some instances in their data in which part A contains New information, and they found that part B of reversed pseudo-clefts more often manifests Given/Inferrable rather than New information (1996: 201).

As for part B of reversed pseudo-clefts opening with *what*, Oberlander & Delin (1996) found that it often expresses presupposed information which is connected to what was said or assumed in the previous discourse (see footnote 7), evoking it and making it relevant again. According to them, in this way reversed pseudo-clefts can function metalingually (cf. Calude 2008: 100-106; Garassino 2014: 71): "This category captures cases in which a speaker is using a presupposition to topicalize, summarize, clarify, or pass judgement on the relevance or importance of a topic under discussion, or the effectiveness of an argument" (Oberlander & Delin 1996: 207).

By far the main discourse function claimed for headless reversed pseudo-clefts, however, is a summative, or ending function (Erdmann 1986; Collins 1991; Oberlander and Delin 1996; Weinert and Miller 1996; Lambrecht 2001; Johansson 2001; Garassino 2014). According to Collins (1991), since this construction is informatively low and has an equational nature, it does not present new content and is thus suitable for summarizing.

We found only one interactional linguistic study dealing with reversed pseudo-clefts – Küttner's article on the English [*that's* + *wh*-clause] format (2020), constructions like *that's what I'm gonna do*. Küttner shows that in third position (Schegloff 1992) the construction functions as a pivot-like nexus between one sequence and the next: It initiates a new sequence and at the same time connects it – via the demonstrative pronoun *that* – to the previous sequence.[8] According to Küttner, the construction sometimes functions to convey agreement (2020: 265, cf. Calude 2008: 102), but more often it functions as an announcement and is usually deployed immediately after an adjacency pair (Schegloff & Sacks 1973), in order to naturally shift from one sequence to the next, thus projecting the continuation of the turn. The format can also be deployed in second position, to initiate actions that were not made relevant by the action of the first pair part. In this way the format is employed as a new first pair part, making a different action relevant. In such cases, there is indeed no alignment with the previous turn, but the coherence of the discourse is preserved: the format serves as a resource for initiating actions that would otherwise seem inappropriate in the sequence and impair its coherence.

[8] Earlier studies (e.g., Weinert & Miller 1996: 193; Oberlander & Delin 1996: 193; Calude 2008: 88–89) have also pointed out that the demonstrative in this structure can function simultaneously in a forward-oriented and backward-oriented fashion.

The headless reversed pseudo-cleft differs in distribution across languages: it is very frequent in spoken English (Oberlander and Delin 1996; Calude 2008), but very rare in Italian (Garassino 2014). In French and Spanish, it does not exist (Lambrecht 2001: 492), and in German and Norwegian syntactically similar structures do exist, but their functions are different: Johansson (2001) found that when an English headless reversed pseudo-cleft is translated into German or Norwegian, other constructions are employed. Thus, to understand the functions and properties of the headless reversed pseudo-cleft in a particular language, a close study of its deployment in the moment-by-moment unfolding of conversation must be conducted.

4. Lexico-semantic, syntactic, and prosodic patterning

4.1. Lexico-semantic patterning

Hopper and Thompson (2008) found that the most frequent predicates in part A of the English pseudo-clefts in their data are the verbs *do* (66%), *happen* (13%), and *say* (8%), showing that in 87% of the cases, part A functions as *a formulaic fragment* classifying upcoming discourse as an action (e.g., *what we're gonna do…*), an event (e.g., what *happened..*), or a paraphrase of what has been said before (e.g., *what I'm saying…*).

In an ongoing cross-linguistic study of pseudo-clefts in French, Hebrew, Swedish and Estonian (Maschler, Pekarek Doehler, Lindström & Keevallik, forthc., cf., Maschler & Pekarek Doehler 2022), we found that the most frequent predicates in the A-parts in our 4 datasets are generally the semantic 'equivalents' of these DO / HAPPEN / SAY verbs, but these account for a much smaller percentage of the tokens compared to the figures in Hopper & Thompson's study (2008), as demonstrated in Table 1.

	English (Hopper & Thompson 2008)	**French**	**Hebrew**	**Swedish**	**Estonian**
DO	66%	9 (16%)	20 (22%)	2 (8%)	2 (14%)
HAPPEN	13%	2 (4%)	20 (22%)	2 (8%)	1 (7%)
SAY	8%	7 (13%)	9 (19%)	3 (13%)	3 (21%)
TOTAL	87%	18 (33%)	49 (53%)	7 (29%)	6 (43%)

Table 1: Lexico-semantic patterning of pseudo-clefts in English, Hebrew, French, Swedish, & Estonian.

An additional 31% of the Hebrew predicates are employed to display a stance concerning what is about to be verbalized in part B (for example, '*what's cool, annoying, scary,…* ' or '*what I like, want…*'). The percentages of stance-taking predicates in pseudo-cleft structures are even greater in French (60%), Swedish (42%), and Estonian (36%) in our databases (Maschler, Pekarek Doehler, Lindström & Keevallik, forthc.).

As for predicates employed in the *ze ma she-* construction, very few of them involve stance-taking predicates (only '*need/ want/ care about*', see Table 2), suggesting that the *ze ma she-* construction is generally *not* employed in our corpus to convey stance.

Predicate	Translation	**N**	**%**
'amar/higid 'say'/ *sha'al* 'ask'/ *hisbir* 'explain'	SAY	22	31.4%
'asa	DO	17	24%
kara 'happen'/*haya 'be'/ yesh 'there is'*	HAPPEN	11	16%
limed	teach	2	3%
hevin	understand	2	3%
tsarix	need	2	3%
ratsa	want	1	1.4%
'ixpat	care about	1	1.4%
kara 'read'; *simen* 'mark'; *hispik* 'accomplish'; *hevi* 'bring'; *nixtav* 'meant to be'; *xava* 'experience'; *hexya* 'revive'; *notsar* 'emerge'; *higi'a* 'arrive'; *tsipa* 'expect'; *ze* 'this'; no predicate		1 each	1.4 % each
TOTAL		70	100%

Table 2: Predicates in reversed pseudo-cleft structures.

Furthermore, whereas by far the great majority of predicates in the pseudo-cleft structure are DO and HAPPEN (22% of all Hebrew predicates in each category, see Table 1), in the reversed pseudo-cleft structure it is the SAY category that is the largest (31.4%), but there are also quite a few DO (24%) and HAPPEN (16%) predicates.[9] However, despite the general great similarity in the most frequent predicates in these two structures, our interactional analysis (Section 5) reveals that the functions of reversed pseudo-clefts are actually quite different from those of pseudo-clefts in Hebrew.

[9] English reversed pseudo-clefts in which part B opens with 'what' exhibit a preference for verbs of thinking, saying, and feeling (Oberlander & Delin 1996: 217). Even prior to 1680, they tended not to manifest the verbs 'do' and 'happen' characteristic of pseudo-clefts (Traugott 2008: 16).

4.2. Syntactic patterning

Another difference between the two constructions is that while part B of pseudo-clefts often spans a discourse unit much longer than a clause (Hopper & Thompson 2008; Günthner 2011; Pekarek Doehler 2011; Maschler & Fishman 2020; Maschler & Pekarek Doehler 2022), part B of reversed pseudo-clefts hardly ever exceeds one clause.[10] This happens in only 3 out of the 70 cases, and in all of them the [*ze ma she-* + clause] structure ends in continuing intonation, followed by a clause explicating the content of *ze* 'that', as in:

(12) *'aval* ***ze*** ***ma*** ***she-****'ani* *'amarti* *la,*
but this.M.SG what that-I say.PST.1.SG to-3F.SG
'but **that's what** I said to her,'

she-ze *lo* */bishvilo/.*
that-it not for-3M.SG
'that it's not /for him/.'

In example (12), the following clause explicates what it is that the speaker 'said to her.'[11]

4.3. Prosodic patterning

Finally, whereas in the case of pseudo-clefts, part B is very often separated by an intonation unit boundary (Chafe 1994; Du Bois 2012) from part A, in the case of our structure, we never find an intonation unit boundary[12] between part A (*ze*) and part

[10] According to Calude (2008: 88), clause complexes have not been attested as ever occurring in Part A of English *non*-demonstrative reversed pseudo-clefts.

[11] Our corpus also manifests 6 tokens in which the [*ze ma she-* + clause] structure ends in sentence-final falling intonation and the following intonation units consist of an increment (Ford, Fox & Thompson 2002) explicating the demonstrative (see, e.g., example (13), l. 20–23) and example (14), l. 11–12).

[12] Intonation unit boundaries are determined in this study, and in the entire Haifa Multimodal Corpus of Spoken Hebrew (Maschler et al. 2024), based on the Santa Barbara discourse transcription method (Chafe 1994, as adapted in Du Bois 2012; Du Bois et al. 1992 and adjusted for Hebrew in Maschler 2017). In this transcription method, we listen for cues indicating intonation unit boundaries, such as initial anacrusis, final lengthening, pitch reset, terminal pitch contour, pauses, initial inbreath, final creaky voice, etc. (for a complete list of features, see Du Bois 2012: 28). Each conversation is transcribed by at least 3 highly experienced transcribers, reaching a high degree of intercoder

B (the *ma she-* part). This provides prosodic support for the claim that *ze ma she-* has crystallized into a construction, a single 'processing chunk' (Bybee 2003: 603).

5. Multimodal interactional analysis

We now turn to the functional aspects of the *ze ma she-* construction. Table 3 shows the functional distribution of the [*ze ma she-*+ clause] structure in our data.[13]

FUNCTION	N	%
Framing prior talk metalingually	29	41.4%
Claim-backing	21	30%
Seeking clarification	4	5.7%
Postulating some 'general truth'	4	5.7%
Disclaiming responsibility	3	4.3%
Summative, ending	2	2.9%
Getting back to previous topic	1	1.4%
Other	6	8.6%
TOTAL	**70**	**100%**

Table 3: Functional distribution of the [***ze ma she-*** + clause] structure.

We see that only two tokens of the construction accomplish a summative, or ending function, the main discourse function claimed for English headless reversed pseudo-clefts (see Section 3 above). In the continuation of this section, we illustrate the remaining functions found for the [*ze ma she-*+ clause] structure in our data, paying particular attention to the position of the structure within the unfolding of actions, to its prosodic features, and to the embodied conduct accompanying it.

5.1. Getting back to a previous topic

We begin with *ze ma she-* as a means for getting back to a previous topic. There is only one such token in our data (1.4% of all cases), but it is an informative one, because it

reliability. Occasionally, we use Praat (Boersma & Weenink 2025), but since visual representation of prosody can be inaccurate, we rely mainly on the hearing of expert transcribers.

[13] The category 'Other' in Table 3 refers to six cases in which we could not pin down the function performed by the structure, thus necessitating further study.

shows that the construction is not necessarily summative, and that it may even carry a projective force, foreshadowing the upcoming action.

In excerpt (13), a couple, Alon and Hillel, are sitting in their living room with their baby, who begins to wake up in Alon's arms. In the moments preceding excerpt (13), following the pseudo-cleft part A *ma she'ani xayav lehaspik maxar* 'what I *must* accomplish tomorrow' (not shown)[14], Alon starts listing several errands he'd like to accomplish: going to Shilav, a baby goods store, preparing the dough for the Challah, the traditional Jewish bread – a list which he does not complete because at this point the baby begins to cry. In the almost two minutes omitted after this in the video clip[15], the couple is debating whether the baby is hungry, Alon goes to the kitchen to prepare a bottle for him, returns, and starts feeding him (not shown). While feeding, Alon addresses the baby:

(13) 'Errands' ('Challah, Shilav, Car' 04:07)

```
1 (Alon):(2.5)'agav,
                by_the_way
          (2.5)'by the way,'

2         ...haxel      me-ha-rega?,
             starting   from-DEF.ART-moment
          '...starting from now?,'

3         ...'ata     muzman    le'exol?,
             you.M.SG invited   eat.INF
          '...you are welcome to eat?,'

4         ...ve-layla   tovi.
             and-night  good_y
          '...and good nighty.'

5         ...'ad    ha-bo*ker.
             until  DEF.ART-morning
        '...until the morning.'
                       *Fig. 1

6         ...'ad    ha-boker.
             until  DEF.ART-morning
          '...until the morning.'
```

Fig. 1

[14] For a detailed analysis of this part of the interaction, see Maschler & Pekarek Doehler 2022.

[15] All video clips analyzed here, with English subtitles, can be accessed at https://drive.google.com/drive/folders/1Y4wV29VhHiaW-3unNIQRUKE0OIWtjnBD

7 Hillel: *@'ad @ha-boker,*
until DEF.ART-morning
'@until the @morning,'

8 *'a*lek.*
as_if
'as if.'
{rolling his eyes}
*Fig. 2

Fig. 2

9 Alon: *'ad ha-boker,*
until DEF.ART-morning
'until the morning,'
{---singing---}

10 Hillel: *yu--!*
PART

11 Alon: *'ad ha-boker,*
until DEF.ART-morning
'until the morning,'
{---singing---}

12 *..'im 'ani 'e/nana/ 'ad ha-bo*
if I until DEF.ART-mor[ning]
'..if I 'e/nana/ until the mor'
{singing, looking to the side, probably at the clock}

13 (inhales)

14 *shmone ve-xamisha.*
eight and-five
'five minutes past eight.'

15 *...'im ba-xayim,*
if in.DEF.ART-life
'...if in life,'
{--singing--}

16 *..'oy 'oy 'oy 'oy.*
{----singing-----}

17 ...(*kisses baby)
*Fig. 3

Fig. 3

Fig. 4

18 (*inhales deeply, straightens back, gazes into space)
*Fig. 4

19 ....*tsk*

→ 20 ..*ze* *ma* *she-'ani* *tsar^i*x.*
this.M.SG what that-I need.PRS.M.SG
'..that's what I need.'
^gaze at Hillel
*Fig. 5

21 ...*xala?,*
'...challah?,'

Fig. 5

22 ...*shilav?,*
'...Shilav?,'

23 ...*sidur* *rexev.*
fixing.of car
'...fixing the car.'

24 ...*'ani* *xaya--v* *lesayem* *'et* *ze.*
I must.PRS.M.SG finish.INF OBJ this.M.SG
'...I must get this done.'

25 ...(inhales through nose)

26 .....*nir'a* *l-i* *re'ali.*
seem.PRS.M.SG to-1SG realistic
'.....seems realistic to me.'

27 Hillel: *shi**lav*** *lo* *yihye* *patuax* *maxar?*
Shilav NEG be.FUT.3M.SG open tomorrow
'Shilav won't be open tomorrow?'
*

28 ..*be-shabat?*
on-Saturday
'..on Saturday?'

Alon is speaking to the baby, expressing his wish that the baby sleep all night until morning (l. 1–6). Hillel ridicules Alon's wish, saying there's no chance this is going to happen (l. 7–8). Alon starts singing and then kisses the baby's forehead (l. 9, 11–17, Fig. 3). At this point Alon takes a deep breath, straightens his back, and focuses his gaze into space (l. 18, Fig. 4). After a rather long pause, focusing his

gaze on Hillel (Fig. 5), he produces a click (l. 19) opening a new sequence (Ben-Moshe & Maschler 2024a),[16] and the *ze ma she-* structure: *ze ma she-'ani tsarix*. – 'that's what I need.' – in sentence-final falling intonation. The primary stress of the intonation unit is on the demonstrative *ze* 'this'. In contrast to what was found in some studies of English reversed pseudo-clefts (e.g., Collins 1991; Oberlander & Delin 1966), in 44% of the *ze ma she-* tokens (32 out of 70), the demonstrative carries the primary stress. In Oberlander & Delin's study, for instance, the figure is 6% (1996: 194).

The construction thus initiates a new sequence, skipping over the almost two-minute baby-episode and tying back to the previous discussion concerning the things Alon must accomplish the next morning, summing them up. However, not only does the construction retroactively frame Alon's previously described actions as things he 'needs to accomplish', it also *projects* an enumeration of those tasks: Challah, Shilav, and fixing the car, an enumeration which indeed follows (lines 21–23). The *ze ma she-* construction thus re-opens a previous discussion, after a very long pause here, and the men continue to discuss those tasks in the following moments (data not shown).

Thus, in contrast to previous literature emphasizing the closing function of English reversed pseudo-clefts, and in line with Küttner's Interactional Linguistic study (2020) (cf. Collins 1991; Oberlander & Delin 1996; Calude 2008 (see footnote 8 above)), we see that the Hebrew construction may carry also a projecting force. However, unlike Küttner's examples, *ze ma she-* also has the capacity to re-open a discussion, skipping over a very long stretch of talk.

5.2. Claim-backing

A claim-backing move is a move made by a participant in order to deal with a dispute (Antaki & Leudar 1990: 284). Our second excerpt demonstrates the second most frequent function of the *ze ma she-* construction in our data – a claim-backing function. Of the 70 tokens in our data, 21 (30%) carry this function. Example (14) comes from a gathering of five friends at Inbal and Omri's flat. Inbal has a stomachache and is therefore resting on the sofa, not visible in the frame. Yair

[16] For a detailed analysis of this part of the interaction, see Ben-Moshe & Maschler 2024a.

suggests preparing a hot water bottle for her to ease the pain but then realizes that they have no kettle. He then thinks of other solutions:

(14) 'Microwave' ('Katamon 2' 00:25:46)

```
1 Ya'ir:'en           lax,
        there_isn't   to-you.F.SG
        'don't you have,'

2       karit   ka-zot,
        pillow  like-this.F.SG
        'this kind of pillow,'

3       she-mitxamemet          be--,
        that-get_warm.PRS.F.SG  in-
        'that can be heated i--n,'
```

Fig. 6

```
4       mikro?
        'a microwave?'

5 Inbal:  lo
          'no.'

6 Ya'ir: ....'efshar        ^lexamem  mayim  ba-mik*ro.
             it's_possible  warm.INF  water  in-the-micro
         '....it's possible to heat up water in the micro.'
                          ^.........---- PUOH -------->1.9
                                                *Fig. 6

7 Inbal:  .../@/

8 Omri:   ....k*en?
              yes
          '....really?'
               *Fig. 7

9 Ya'ir:  ^...^m m--
          ...m m--
          ^moves PUOH to the left, shakes head^---holds PUOH---> 1.10
```

Fig. 7

```
→ 10      ze          ma    she-hu   yode'a          la'a*sot.^
          this.M.SG   what  that-he  know.PRS.M.SG   do.INF
          'that's what it can do.^'
                                                *Fig. 8
```

Fig. 8

11 *ze* *kol* *ma* *she-mikro* *yode'a* *la'asot.*
this.M.SG every what that-micro know.PRS.M.SG do.INF
'that's all a microwave can do.'

12 *lexamem* *mayim.*
warm.INF water
'heat up water.'

13 Omri: */ha-'inya/*
the-thin[g]
'/the thing/'

14 *hu martit* *molekulot* *mayim.*
he vibrate.PRS.M.SG molecule.of.PL water
'it vibrates water molecules.'

15 *'ani* *lo* *yode'a* *'im* *hu* *mexamem* *'otam*
I NEG know.PRS.M.SG if he warm.PRS.M.SG OBJ-3m.pl
lenekudat *retixa.*
to-point.of boiling
'I don't know if it heats them up to a boiling point.'

16 Ya'ir: *..hu* *mexamem*
he warm.PRS.M.SG
'..it heats'

17 *...hu* **mexamem* *'otam.*
he warm.PRS.M.SG OBJ-3m.PL
'...it heats them up.'
*Fig. 9

Fig. 9

Upon realizing that they have no kettle, Yair asks Inbal whether she has a pillow that could be heated up in the microwave so that she could put it on her stomach to relieve the pain (l. 1–4). Following her negative reply (l. 5), Yair suggests, while holding his left hand in a Palm Up Open Hand gesture (PUOH), that they could heat up water in the microwave (l. 6, Fig. 6). Yair keeps holding his hand in this gesture until Omri, after a rather long pause, tilts his head, raises his eyebrows, and produces a doubtful *ken?* 'really?' in rising intonation (l. 8, Fig. 7). Yair then moves his hand, still in the PUOH, horizontally to the left, while shaking his head, and produces a response to Omri's doubt: *ze ma she-hu yode'a la'asot.* 'that's what it can do'. (l. 9–10, Fig. 8).

Many of our claim-backing tokens – 56% of them (10 of the 18 in which the speaker is visible) – are accompanied by the PUOH, whereas none of the tokens of the construction accomplishing other functions are. The PUOH has been associated in various studies with obviousness and shared knowledge (e.g., Kendon 2004; Müller 2004; Cooperrider et al. 2018; Maresse et al. 2021; Inbar & Maschler 2023).

The structure – with the verb 'to do' – points out the function of the microwave (heating up water) in order to present evidence that Yair's suggestion was indeed sensible. According to Antaki and Leudar, ''A backing move does two things – it marks the move as disputable in a particular way, and at the same time it presents grounds to deal with that disputability" (Antaki & Leudar 1990: 284). In other words, the [*ze ma she-*+ clause] structure functions here as 'claim-backing'. Accompanying it with the PUOH, which functions as an epistemic stance marker (Inbar & Maschler 2023), upgrades the backing move by marking the claim as obvious, shared knowledge.

Indeed, right after producing the structure, in final falling intonation, Yair produces an upgraded version of it: *ze <u>kol</u> ma she-mikro yode'a la'asot.* 'that's <u>all</u> a microwave can do.' (l. 11), further backing up his claim. He then increments his utterance with the infinitive phrase 'to heat up water.' (l. 12), specifying the demonstrative pronoun employed in the construction. Following the increment, Omri continues to argue with a contrasting assertion (l. 13-15), which Yair counters as well (l. 16-17, Fig. 9).

In this token of our construction, there are two primary stresses – one on *ze* 'this' and one on the final syllable of *la'asot* 'to do'. This can clearly be seen in the following Praat diagram[17] (Diagram 1):

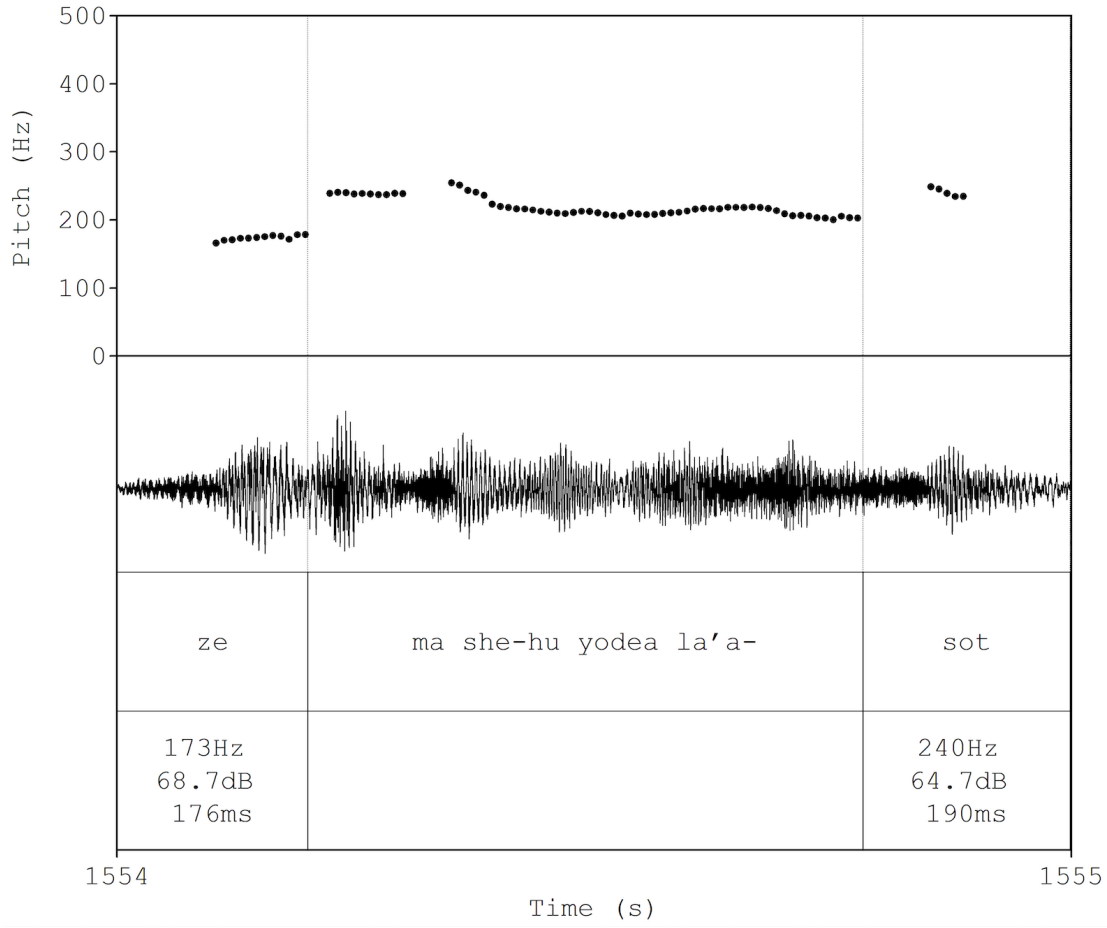


Diagram 1: Praat diagram of *ze ma she-hu yode'a la'asot.* 'that's what it can do'. (l. 10)

Diagram 1 shows that both syllables, *ze* 'this' and *sot* of *la'asot* 'to do', are of roughly the same length: 176 and 190 milli-seconds. The syllable *ze* is somewhat lower in pitch than *sot* (173 Hertz vs. 240 Hertz), but *ze* is higher in intensity than *sot* (68.7 Decibels vs. 64.7 Decibels).[18] These features all contribute to our conceiving both syllables as stressed to the same degree.

[17] We are indebted to Nadav Matalon for help with this acoustic analysis.

[18] As one of the reviewers noted, measurements can be notoriously unreliable, especially in conversational data not collected under controlled laboratory conditions. This diagram is therefore provided here as secondary support for our claim. The main support comes from the fact that at least six different expert transcribers have listened to this intonation unit, all agreeing that it is not possible to determine which of the two stresses of intonation unit 10 is more prominent.

Eighteen of the 70 tokens in our data, 26%, manifest two primary stresses - one on the demonstrative, the other on some syllable in the clause. This is reminiscent of Weinert & Miller's finding that "[t]he large majority of reversed *wh*-clefts have a stressed element in both the clefted constituent and the cleft clause" (1996: 189) (cf. Lambrecht 2001: 482).

5.3. Framing prior talk metalingually

Example (15) illustrates the most frequent function of the [*ze ma she-* + clause] structure in our data – framing prior talk metalingually (cf. Oberlander & Delin 1996; Calude 2008). Of the 70 tokens in our data, 29 (41.4%) carry this function. This example involves one of the only stance-taking predicates found in this structure in our data. In this interaction, a conversation between two close friends, Kelsey recommends the Israeli TV series 'When Heroes Fly' to her friend Naomi:

(15) 'A Brilliant TV Series' ('Armon Hanatsiv' 01:06:05)

```
1 Kelsey: ra'iti,
          see.PST.1SG
         'I watched,'

2      ...male  prakim     shel  bishvila  giborim 'afim,
          full  chapter.PL of    for-3F.SG  hero.PL  fly.PRS.M.PL
       '...lots of episodes of 'When Heroes Fly','

3 Naomi: (nodding and yawning)

4 Kelsey:                ve-siyamti        'et kol  ha-sidra
                         and-finish.PST.1SG OBJ all  DEF.ART-series
                         'etmol?,
                         yesterday
                        'and I finished the whole series yesterday?,'

5        ...bimkom       lilmod?,
            instead_of   study.INF
         '...instead of studying?,'
```

6 Naomi: (smiling with puckered lips)

7 Kelsey: *way,*
PART,

8 ...*'at* *xayevet* *lir'ot* *'et* *ze.*
you.F.SG must.PRS.F.SG see.INF OBJ this.M.SG
'...you've got to watch it.'

9 Naomi: *sidra* *me'ula.*
series superb
'[it's] a brilliant series.'

Fig. 10

10 ..*'ani* *ra'iti.*
I see.PST.1SG
'..I've watched it.'

11 Kelsey: ..^*sidra* *me'u*la,*^
series superb
'..a brilliant series,'
^right hand IF touches left hand IF in listing gesture^
*Fig. 10

Fig. 11

12 ^*ha-sax*kanim* *xa^tixim,*
DEF.ART-actor.PL handsome.PL
'the actors are hunks,'
^moves right hand IF to left hand MF^
^presses right hand
IF against left MF,
pushing it backwards ->1.16
*Fig. 11

13 Naomi: *'ani* *ra'iti* *kshe-hi* *hayta* *ba-televizya*
I see.PST.1SG when-she be.PST.3F.SG in-DEF.ART-TV
'I watched it when it was on TV'

14 Kelsey:...*kol* *ka--x,*
{*smiling*}
all this_way
'...su--ch [hunks],'

15 Naomi: ..@@@@@

→ 16 Kelsey: *ze* *ma^* *#she-h*aya* *li* *'ixpat.#*
{----------*smiling*----------}
this.M.SG what that-be.PST.3M.SG to-1SG care
'that's what I cared about.'
#detaches IF from left hand, switches into RIF gesture#

*Fig. 12

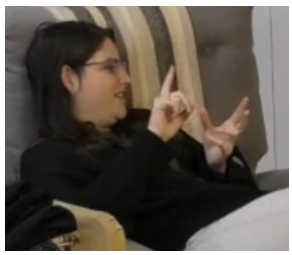


Fig. 12

17 Naomi: @@@@
{*ff*}

18 Kelsey: *^be-'emet,^*
in-truth
'really,'
^brings right
IF closer to left MF and touches it^

19 *'ani kaze,*
I like-this.M.SG
'I'm like,'

20 ...(inhales through teeth)

21 Naomi: @--

22 Kelsey: .../'o/.

23 Naomi: ...'e--,
...u--h,

24 *Matan* *ra'a* *ta-sidra?*
Matan see.PST.3M.SG OBJ.DEF.ART-series
'did Matan watch the series?'

25 Kelsey: ...ken.
'...yes.'

Kelsey evaluates the TV series, detailing a list of features characterizing the series. First on her list, she agrees with Naomi and repeats her evaluation, '...a brilliant series,' (l. 11, Fig. 10). Second on her list is the evaluation that the actors are *xatixim*

'hunks' (l. 12, Fig. 11). She then upgrades this evaluation by adding the lengthened intensifier *kol ka--x* 'su--ch [hunks]' (l. 14), while smiling.

At this point she shifts from evaluating the series and listing its great features, to commenting on her own evaluation. With a bit of self-irony (marked by her smile), looking back at her evaluation and still smiling, she produces the metalingual comment: *ze ma she-haya li 'ixpat.* '**that's** what I cared about.' (l. 16, Fig. 12), with the primary stress of the intonation unit on the demonstrative *ze* 'that', as in example (13). She employs the '[*ze ma she-* + clause]' structure to *retroactively frame* the second item on her list and evaluate it. Thus, the *ze ma she-* construction marks here a shift in the discourse from narrating about the world to metalingual commentary.

This shift is marked also by the speaker's embodied conduct: when listing the features of the series, she deploys listing gestures (Inbar 2020): she moves her right-hand index finger from the index finger to the middle finger of her left hand, counting the features with her fingers (Fig. 10, 11). However, when shifting to metalingual commentary, she detaches her hands from each other and performs the raised index finger gesture (RIF, Fig. 12).

Framing prior talk metalingually via this structure is done in order to accomplish a variety of actions in our data, in the case of example (15) – to ironically evaluate one's own utterance. It is often accompanied by some pointing gesture – here the raised index finger gesture marking contrariness to expectations (Inbar 2022) – *ze ma she-haya li 'ixpat.* '**that's** what I cared about.' (l. 16), as opposed to the content of the series, for example. Thus, the prosodic features of the utterance (primary stress on *ze*) and the embodied conduct of the speaker are perfectly coordinated in production of the [*ze ma she-* + clause] structure. This constitutes embodied support for Chafe's observation that the primary stress of an intonation unit may fall on a pronoun for the expression of contrastiveness (Chafe 1994: 76–78).

5.4. Disclaiming responsibility

Our final excerpt manifests two tokens of the structure, each implementing a different action. The first token is employed for disclaiming responsibility and manifests the predicate *'amar* 'say', the most frequent predicate in our construction (see Table 2). This interaction comes from a family dinner, in which the mother tells the father about an earlier conversation she had had with their two children, during which

Assaf, the younger one, mentioned that this year nobody in kindergarten had talked to them about Prime Minister Rabin's assassination. In response, the father asks his son:

(16) 'Peace' ('Rabin's Murder' 00:15)

```
1 Father: shana    she-'avra            dibru?
          year     that-PAST.PST.3F.SG  talk.PST.3M.PL
          'did they talk to you [about it] last year?'

2 Assaf:  ...ken,
          '...yes,'

3         ...nir'a        li      she--
             seem.PRS.M.SG to-1SG  that
          '...I think tha--t'

4 Ron:          kol   ma    she-hu    zaxar,
                all   what  that-he   remember.PST.3M.SG
                'all he remembered,'

5         ...ze         she-rabin     haya          'ish,
             this.M.SG  that-Rabin    be.PST.3M.SG   man
          '...was that Rabin was a man,'

6         she--,
          that
          'who--,'

7         ...kol  ha-zman     /hu/  'asa          shalom.
             all  DEF.ART-time  he    do.PST.3M.SG  peace
          '...all the time /he/ made peace.'

8 Father: ...kol @ha-zman     @'asa          sha*lom?
             all DEF.ART-time  do.PST.3M.SG   peace
          '...all the time made peace?'
                                              *Fig. 13
9 Assaf:                      'e--,
                              u--h,

10        'a--h!
           o--h!

→ 11 Ron:  ze        ma    she-*'assaf  'amar!
           this.M.SG what  that-Assaf   say.PST.3M.SG
           'that's what Assaf said!'
                                *Fig. 14
```

Fig. 13

Fig. 14

→ 12 Father:....*ze* *ma* *she-limdu* *'etxem?*
this.M.SG what that-teach.PST.3M.PL OBJ-2M.PL
'....that's what they taught you?'

13 Mother: *lo--,*
'no--,'

14 *hu 'amar* *she*
he say.PST.3M.SG that
'he said that'

15 *...kol* *ha-zman* *rak* *ratsa* *shalom.*
all DEF.ART-time just want.PST.3M.SG peace
'...he all the time just wanted peace.'

16 Ron: *naxon.*
'right.'

Ron, the older brother, says that all Assaf remembered, concerning what he was told in kindergarten about Rabin's assassination, was "that Rabin was a man, who--, ...all the time made peace." (l. 5-7). The father ridicules this assertion by repeating it with a smile, slight laugher, and a primary stress on *kol* 'all': '...all the time made peace?' (l. 8, Fig. 13). Ron then disclaims responsibility for the utterance by announcing: *ze ma she-'assaf 'amar!* 'that's what Assaf said!' (l. 11, Fig. 14) with the primary stress of the intonation unit on 'Assaf', the responsible agent of the action, and while pointing at him with his left hand (Fig. 14).

There are only three tokens performing this function throughout our data (4.3% of all cases), but they all manifest only one primary stress, which falls on the 'responsible agent', i.e., in the clause.

5.5. Seeking clarification

The last function illustrated here, also not that common in our data (4 tokens, 5.7%) consists of seeking clarification. One token is found in the continuation of example (16). Following Ron's disclaiming responsibility (l. 11), the father seeks clarification and at the same time criticizes the kindergarten staff via the *ze ma she-* construction in sentence-final rising intonation: *...ze ma she-limdu 'etxem?* '...that's what they taught you?' (l. 12). This time, the primary stress is on the predicate of the *ma she-* clause – 'taught', the object of the criticism, conveying the father's disapproval of the teaching at kindergarten. The mother then clarifies that Assaf had said that Rabin all the time just *ratsa* 'wanted' peace, rather than all the time just *'asa* 'made' peace (l. 13-15).

All four tokens performing the seeking clarification function are verbalized in final rising intonation – by far the most frequent prosodic pattern for Hebrew requests for confirmation (Ben-Moshe & Maschler 2024b) and for Hebrew interrogatives in general (Ozerov 2019). In 3 out of the 4 cases, the primary stress is in the clause (rather than on the demonstrative).

5.6. Postulating some 'general truth'

Four tokens throughout the database (5.7%) appear in short utterances that have sedimented in the language for postulating some 'general truth' about some frequently encountered situation, such as *ze ma she-yesh* 'that's what there is'[19], *ze ma she-ze* 'that's what it is', *ze ma she'osim* 'that's what one does'. We do not illustrate them here for lack of space.

[19] Further support for the crystallization of this utterance comes from the fact that a morphosyntactically reduced version of it exists, although not in our data: *ze ma yesh* 'that's what there is', without the complementizer *she-* preceding the clause.

6. Summary and conclusion

The Hebrew *ze ma she-* construction manifests several properties suggesting that it has crystallized as a single 'processing chunk' (Bybee 2003: 603). Prosodically, we have seen that the demonstrative *ze* in this construction (part A) is never separated from the *ma she-* segment (part B) by an intonation unit boundary. This is initial evidence against viewing parts A and B as two sides of an A COP B equational structure, as claimed in previous studies of reversed pseudo-clefts in a variety of languages (see Section 3). Furthermore, syntactically, unlike pseudo-clefts, we have seen that in Hebrew, a language allowing non-copular nominal clauses, no copula is ever found separating parts A and B of the [*ze ma she-* + clause] structure. Both prosody and syntax thus provide evidence against viewing this structure as a 'reversed' form of the Hebrew pseudo-cleft, in which the copula often occurs, and parts A and B tend not to be verbalized within the same intonation unit (Maschler & Fishman 2020; Maschler & Pekarek Doehler 2022).

Further support for the sedimentation of the *ze ma she-* construction comes from examining its lexical properties: As shown in Table 2, 71.4% of all tokens of the [*ze ma she-* + clause] structure manifest one of three verbs – SAY, DO, or HAPPEN – with SAY accounting for approximately a third of all tokens of the structure throughout the data. The structure is thus relatively fixed in terms of the predicates it occurs with, further pointing towards crystallization and suggesting a strong metalingual component in its usage. Indeed, our functional analysis revealed that 41.4% of all tokens function to frame prior talk metalingually (Table 3, Section 5.3).

The paucity of stance-taking predicates in this construction (only 5.8% of all tokens, Table 3) suggests that unlike pseudo-clefts, the Hebrew [*ze ma she-* + clause] structure has little to do with displaying stance. This contributes further evidence against viewing the structure as a 'reversed' form of the pseudo-cleft, in which nearly a third of all predicates constitute stance-taking predicates (Maschler & Fishman 2020; Maschler & Pekarek Doehler 2022).

Additional attestation for the crystallization of the construction comes from its systematic deployment throughout the corpus. We have seen that it is consistently employed for two main purposes: framing prior talk metalingually (41.4% of all tokens, Section 5.3), and claim-backing (30%, Section 5.2). Thus, 71.4% of all tokens accomplish one of two actions in interaction. The remaining tokens function in several less frequent functions: seeking clarification (5.7%, Section 5.5), postulating some

general truth (5.7%, Section 5.6), disclaiming responsibility (4.3%, Section 5.4), and getting back to a previous topic (1.4%, Section 5.1). Only two tokens throughout the data (2.9%) display the summative function claimed as by far the most common function for English reversed pseudo-clefts (Erdmann 1986; Collins 1991; Weinert & Miller 1996; Oberlander & Delin 1996; Lambrecht 2001; Johansson 2001; Garassino 2014), showing that functionally, the Hebrew [*ze ma she-* + clause] structure has little in common with what has been claimed for English reversed pseudo-clefts.

The fact that some of the Hebrew [*ze ma she-* + clause] tokens consist of utterances that have become fixed expressions for postulating some 'general truth', such as *ze ma (she-)yesh* 'that's what there is', *ze ma she'osim* 'that's what one does', *ze ma she-ze* 'that's what it is' (Section 5.6) constitutes further evidence of the crystallization of the structure. With certain highly fixed predicates, such as the frozen Hebrew existential *yesh* 'there is/are' (cf. Auer & Maschler 2013: 157–159), the impersonal form of the verb with the widest semantic scope, *'asa* 'do' (cf. Polak-Yitzhaki 2017), or with no predicate at all (in the case of *ze ma she-ze* 'that's what it is'), additional sedimentation has occurred, such that entire coined phrases have emerged to deal with frequently-encountered situations.

Additional evidence for the crystallization of the construction comes when examining the speaker's embodied conduct. We have seen that the claim-backing function is accompanied by the PUOH in 56% of its occurrences, whereas none of the other functions is accompanied by this gesture. The co-occurrence of a particular function of the construction with a particular type of embodied conduct is further evidence of crystallization, one involving not only language but also the speaker's bodily conduct.

We have found some correlation between prosody and the function of the [*ze ma she-* + clause] structure. All seeking-clarification tokens were verbalized in sentence-final rising intonation, and all disclaiming-responsibility tokens manifest the primary stress of the intonation unit on the responsible agent.[20] Other than this, we found no correlation between the position of the primary stress of the intonation unit and the function of the [*ze ma she-*+ clause] structure. Furthermore, 26% of all tokens were accompanied by two primary stresses – one on the demonstrative, the other on some element in the clause (cf. Weinert & Miller 1996: 189). In contrast to what has been

[20] However, our numbers are small for these two categories. Furthermore, qualitative analysis of a significantly larger number of tokens of the [*ze ma she-* + clause] structure may reveal additional correlations between function, lexico-semantic variation, prosody, and embodied conduct.

claimed in most studies of English reversed pseudo-clefts (see Section 3), we have found that the demonstrative very often (in 44% of all cases) carries the primary stress of the intonation unit. This is in stark contrast to Information Structure approaches, which claim that demonstratives constitute Given information and are therefore generally unstressed[21] (see Section 3). Stress placement, we argue, is a function of several variables, such as New information, contrastiveness, and the general point being made by the utterance in the particular context. Moreover, in the case of our structure, the stress is often accompanied by prominent embodied conduct, such as pointing, further supporting a holistic view of grammar, involving both acoustic and visual behavior.

In conclusion, and in the context of this special issue focusing on naturally occurring data in and beyond linguistic typology, we hope to have shown that it is not sufficient to focus on naturally occurring data. One must study this type of data employing a specific methodology and within a particular theoretical framework. With the exception of Küttner's Interactional Linguistic study, previous studies of English reversed pseudo-clefts, although often based on spoken corpora, did not examine the data paying close attention to the temporality and contingency of the moment-by-moment incremental unfolding of interaction. Once such an approach is taken, claims of previous studies are not necessarily borne out by naturally occurring spoken data. Similarly, our findings do not support the traditional Hebrew grammatical analysis of the [*ze ma she-* + clause] structure as consisting of a nominalized clause functioning as predicate, embedded in the matrix clause. Nor do they support the traditional grammatical analysis of this construction as constituting a 'reversed' version of a Hebrew pseudo-cleft (Azar 1992). Finally, to the best of our knowledge, no previous study has investigated the embodied conduct accompanying employment of reversed pseudo-clefts, in any language. Our multimodal interactional analysis thus sheds new light on the construction. It illuminates the interlaced nature of grammar, the body, and interaction: Syntax, lexicon, prosody, and the body are resources that "mutually elaborate each other to create a whole that is both greater than, and different, from any of its constituent parts" (Streeck et al. 2011: 2).

[21] One clear exception to this is the occurrence of contrastive information (Chafe 1994: 76–78), as noted in Section 5.3.

Acknowledgements

We are very grateful to Caterina Mauri for having organized the highly stimulating conference 'Naturally Occurring Data in and beyond Linguistic Typology' during May 2023 in Bologna and for the invitation to participate in it. Another version of this study was presented at the 'Complex Syntax for Interaction' panel at the 18th International Pragmatics Association (IPrA) Conference in Brussels in June 2023. We thank all the participants for their important questions and comments. Finally, we are grateful to two anonymous reviewers.

Abbreviations

1 = 1st person	FUT = future	PART = particle
3 = 3rd person	INF = infinitive	PL = plural
ACC = accusative marker	M = masculine	PRS = present
DEF.ART = definite article	NEG = negation marker	PST = past
F = feminine	OBJ = object	SG = singular

References

Antaki, Charles & Ivan Leudar. 1990. Claim-backing and other explanatory genres in talk. *Journal of Language and Social Psychology*, 9. 279–292.

Auer, Peter. 2005. Projection in interaction and projection in grammar. *Text,* 25. 7-36.

Auer, Peter. 2009. On-line syntax: Thoughts on the temporality of spoken language. *Language Sciences,* 31. 1-13.

Auer, Peter & Yael Maschler. 2013. Discourse or Grammar? VS patterns in spoken Hebrew and spoken German narratives. *Language Sciences,* 37. 147–181.

Azar, Moshe. 1992. *Likrat havanat mivne hamishpat hamemukad ba'ivrit bat yameynu* ['Towards an understanding of the structure of the focused sentence in contemporary Hebrew']. In Ouzzi Ornan, Rina Ben Shachar & Gideon Touri (eds.), *Ha'ivrit safa xaya* ['Hebrew - A Living Language'], 87–99. Haifa: University of Haifa Press.

Ball, Catherine N. 1991. *The historical development of the it-cleft.* Philadelphia: University of Pennsylvania. Doctoral Dissertation.

Ben-Moshe, Yotam & Yael Maschler. 2024a. Hebrew clicks: From the periphery of language to the heart of grammar. *Journal of Pragmatics*, 229. 19–39.

Ben-Moshe, Yotam & Yael Maschler. 2024b. Requests for confirmation sequences in Hebrew. *Open Linguistics,* 10(1). 20240028. https://doi.org/10.1515/opli-2024-0028

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Boersma, Paul & David Weenink. 2025. Praat: Doing phonetics by computer. https://www.fon.hum.uva.nl/praat/ [Computer software]. Version 6.4.47.

Bybee, Joan. 2003. Mechanisms of change in grammaticization: The role of frequency. In Brian D. Joseph & Richard D. Janda (eds.), *The Handbook of Historical Linguistics,* 602–623. Oxford: Blackwell.

Calude, Andreea S. 2008. Demonstrative clefts and double cleft constructions in spontaneous spoken English. *Studia Linguistica*, 62(1). 78–118.

Chafe, Wallace. 1994. *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. Chicago: University of Chicago Press.

Collins, Peter C. 1991. *Cleft and Pseudo-Cleft Constructions in English*. London and New York: Routledge.

Cooperrider, Kensy, Natasha Abner & Susan Goldin-Meadow. 2018. The palm-up puzzle: Meanings and origins of a widespread form in gesture and sign. *Frontiers in Communication*, 3. https://doi.org/10.3389/fcomm.2018.00023

Couper-Kuhlen, Elizabeth & Margaret Selting. 2018. *Interactional Linguistics: Studying Language in Social Interaction*. Cambridge: Cambridge University Press.

De Cesare, Anna-Maria. 2014. *Frequency, Forms and Functions of Cleft Constructions in Romance and Germanic: Contrastive, Corpus-Based Studies*. Berlin, München, Boston: De Gruyter Mouton.

Du Bois, John W. 2012. *Representing Discourse*. Unpublished manuscript, Linguistics Department, University of California at Santa Barbara (Fall 2012 version). https://www.linguistics.ucsb.edu/research/santa-barbara-corpus.

Du Bois, John W. 2014. Towards a dialogic syntax. *Cognitive Linguistics,* 25(3). 359–410.

Du Bois, John W., Susanna Cumming, Stephan Schuetze-Coburn & Paolino Danae. 1992. Discourse Transcription. *Santa Barbara Papers in Linguistics*, vol. 4. Santa Barbara: Department of Linguistics, University of California, Santa Barbara.

Erdmann, Peter. 1986. A note on reverse wh-clefts in English. In Dieter Kastovsky & Aleksander Szwedek (eds.) *Linguistics across Historical and Geographical Boundaries: Vol 1: Linguistic Theory and Historical Linguistics. Vol 2: Descriptive, Contrastive, and Applied Linguistics. In Honour of Jacek Fisiak on the Occasion of His Fiftieth Birthday*, 851-858. Berlin, New York: De Gruyter Mouton.

Ford, Cecilia E., Barbara A. Fox & Sandra A. Thompson. 2002. Constituency and the grammar of turn increments. In Cecilia E. Ford, Barbara A. Fox & Sandra A. Thompson (eds.), *The Language of Turn and Sequence*,14–38. Oxford: Oxford University Press.

Garassino, Davide. 2014. Reverse Pseudo-cleft sentences in Italian and English: A contrastive analysis. *Tra romanistica e germanistica: lingua, testo, cognizione e cultura/Between Romance and Germanic: language, text, cognition and culture*, 55-74.

Geluykens, Ronald. 1984. *Focus phenomena in English: An empirical investigation into cleft and pseudo-cleft sentences.* (Tech. Rep. No. 36). Antwerp: Universitaire Installing Antwerpen, Departement Germaanse.

Geluykens, Ronald. 1988. Five types of clefting in English discourse. *Linguistics*, 26. 823–841.

Goffman, Erving. 1981. *Forms of Talk*. Philadelphia: University of Pennsylvania Press.

Goodwin, Charles. 2000. Action and embodiment within situated human interaction. *Journal of Pragmatics*, 32(10). 1489-1522.

Goodwin, Charles. 2018. *Co-Operative Action*. Cambridge: Cambridge University Press.

Günthner, Susanne. 2011. Between emergence and sedimentation: Projecting constructions in German interactions. In Peter Auer & Stephan Pfänder (eds.), *Constructions: Emerging and Emergent*, 156-185. Berlin: Walter de Gruyter.

Günthner, Susanne & Paul J. Hopper. 2010. Zeitlichkeit & sprachliche Strukturen: Pseudo-clefts im Englischen und Deutschen. *Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion* 11. 1–28.

Hopper, Paul J. 1987. Emergent grammar. In Jon Aske, Natasha Beery, Laura Michaelis, & Hana Filip (eds.), *Proceedings of the thirteenth annual meeting of the Berkeley Linguistics Society* 13, 139-157. Berkeley: Berkeley Linguistics Society.

Hopper, Paul J. 2001. Grammatical constructions and their discourse origins: Prototype or family resemblance? In Martin Pütz, Susanne Neimeier & René Dirven (eds.), *Applied Cognitive Linguistics I: Theory and Language Acquisition*, 109-129. Berlin / New York: Mouton de Gruyter.

Hopper, Paul J. 2004. The openness of grammatical constructions. *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, 40(2). 153-175.

Hopper, Paul J. 2011. Emergent grammar and temporality in interactional linguistics. In Peter Auer & Stefan Pfänder (eds.), *Constructions: Emerging and Emergent*, 22–44. Berlin: Walter de Gruyter.

Hopper, Paul J. 2020. Afterword. In Yael Maschler, Simona Pekarek Doehler, Jan Lindström, & Leelo Keevallik (eds.), *Emergent Syntax for Conversation: Clausal Patterns and the Organization of Action*, 331–338. Amsterdam/Philadelphia: John Benjamins.

Hopper, Paul J. & Sandra A. Thompson. 2008. Projectability and clause combining in interaction. In Laury Ritva (ed.), *Crosslinguistic Studies of Clause Combining: The Multifunctionality of Conjunctions,* 99-123. Amsterdam: John Benjamins.

Inbar, Anna. 2020. *'al tafkidey hareshimot basiax: nituax hamexvot hanilvot lereshimot basiax ha'ivri hadavur* ['On the functions of lists in discourse: The analysis of gestures coordinated with list constructions in spoken Israeli Hebrew']. *divrey haxug hayisre'eli levalshanut* 22 [*Proceedings of the 33rd–35th Annual Meetings of the Haiim Rosén Israeli Linguistic Society* 22]. 69–84.

Inbar, Anna. 2022. The raised index finger gesture in Hebrew multimodal interaction. *Gesture*, 21(2/3). 264–295.

Inbar, Anna & Yael Maschler. 2023. Shared knowledge as an account for disaffiliative moves: Hebrew *ki* 'because'-clauses accompanied by the Palm-Up Open-Hand Gesture. *Research on Language and Social Interaction,* 56(2). 141-164, DOI: 10.1080/08351813.2023.2205302.

Jespersen, Otto. 1949. *A Modern English Grammar on Historical Principles*. Part VII. Copenhagen: E. Munksgaard.

Johansson, Stig. 2001. The German and Norwegian correspondences to the English construction type *that's what*. *Linguistics,* 39(3). 583-605.

Kendon, Adam. 2004. *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press.

Küttner, Uwe-A. 2020. Tying sequences together with the [*That's* + *Wh*-Clause] format: On (Retro-) sequential junctures in conversation. *Research on Language and Social Interaction*, 53(2). 247–270. https://doi.org/10.1080/08351813.2020.1739422

Lambrecht, Knut. 2001. A framework for the analysis of cleft constructions. *Linguistics*, 39(3). 463–516.

Linell, Per. 2009. *Rethinking Language, Mind, and World Dialogically: Interactional and Contextual Theories of Human Sensemaking*. Information Age Publishing.

Marrese, Olivia H., Chase W. Raymond, Barbara A. Fox, Cecilia E. Ford & Megan Pielke. 2021. The grammar of obviousness: The Palm-Up gesture in argument sequences. *Frontiers in Communication*, 6, 663067. https://doi.org/10.3389/fcomm.2021.663067

Maschler, Yael. 2017. The emergence of Hebrew *loydea/loydat* ('I dunno MASC/FEM') from interaction: Blurring the boundaries between discourse marker, pragmatic marker, and modal particle. In Andrea Sansò & Chiara Fedriani (eds.), *Pragmatic Markers, Discourse Markers and Modal Particles: New Perspectives*, 37–69. Amsterdam/Philadelphia: John Benjamins.

Maschler, Yael & Stav Fishman. 2020. From multi-clausality to discourse markerhood: The Hebrew *ma she-* 'what that' construction in pseudo-cleft-like structures. *Journal of Pragmatics*, 159. 73–97.

Maschler, Yael, Jan Lindström, & Elwys De Stefani. 2023. Pseudo-clefts: An interactional analysis across languages. In Elwys De Stefani, Jan Lindström & Yael Maschler (eds.), *Pseudo-Clefts from a Comparative Pragmatic Typological Perspective*. Special issue of *Lingua*, 291. Article 103538.

Maschler, Yael & Simona Pekarek Doehler. 2022. Pseudo-cleft-like structures in Hebrew and French conversation: The syntax-lexicon-body interface. *Lingua* 280. Article 103397.

Maschler, Yael, Simona Pekarek Doehler, Jan Lindström & Leelo Keevallik. (forthc.). *The grammar-body interface: A cross-linguistic analysis of pseudo-cleft-like constructions in Hebrew, French, Swedish and Estonian interaction*.

Maschler, Yael, Hilla Polak-Yitzhaki, Galith Aghion, Ophir Fofliger, Nikolaus Wildner, Yotam M. Ben Moshe, Rotem Lagil, Shira Saar, Anna Inbar & Yuval Geva. 2024. *The Haifa Multimodal Corpus of Spoken Hebrew*. https://cris.haifa.ac.il/en/publications/the-haifa-multimodal-corpus-of-spoken-hebrew/

Mondada, Lorenza. 2006. Challenges of multimodality: Language and body in social interaction. *Journal of Sociolinguistics*, 20(3). 336–366.

Mondada, Lorenza. 2019. Conventions for multimodal transcription. https://www.lorenzamondada.net/multimodal-transcription.

Müller, Cornelia. 2004. Forms and uses of the palm up open hand: A case of a gesture family? In Cornelia Müller & Roland Posner (eds.), *The semantics and pragmatics of everyday gestures*, 233–256. Berlin: Weidler.

Oberlander, Jon & Judy Delin. 1996. The function and interpretation of reverse wh-clefts in spoken Discourse. *Language and Speech*, 39(2-3). 185–227.

Ozerov, Pavel. 2019. This is not an Interrogative: The prosody of "wh-questions" in Hebrew and the sources of their questioning and rhetorical interpretations. *Language Sciences*, 72. 13–35. doi: 10.1016/j.langsci.2018.12.004.

Pekarek Doehler, Simona. 2011. Clause-combining and the sequencing of actions: Projector constructions in French talk-in-interaction. In Laury Ritva & Suzuki Ryoko (eds.), *Subordination in Conversation: A Cross-Linguistic Perspective*, 103-148. Amsterdam: John Benjamins.

Polak-Yitzhaki, Hilla. 2017. *ben dibur le'asiya: hapo'al* 'asa '*ufe'alim 'axerim basiax ha'ivri hadavur* ['Between Saying and Doing: The Verb *'asa* ('do') and Other Verbs in Spoken Hebrew Discourse']. Haifa: University of Haifa. Doctoral dissertation.

Prince, Ellen. 1978. A comparison of WH-clefts and IT-clefts in English discourse. *Language*, 54. 883-906.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Saussure, Ferdinand de. 1959 [1913]. *Course in General Linguistics*. Edited by Charles Bally and Albert Sechehaye, in collaboration with Albert Reidlinger. Translated from French by Wade Baskin. New York: McGraw-Hill.

Schegloff, Emanuel A. 1992. Repair after next turn: The last structurally provided for place for the defense of intersubjectivity in conversation. *American Journal of Sociology,* 95(5). 1295–1345.

Schegloff, Emanuel A. & Harvey Sacks. 1973. Opening up closings. *Semiotica,* 8. 289–327.

Streeck, Jürgen, Charles Goodwin & Curtis LeBaron (eds.). 2011. *Embodied Interaction: Language and Body in the Material World*. Cambridge: Cambridge University Press.

Stubbs, Michael. 1983. *Discourse Analysis: The Sociolinguistic Analysis of Natural Language*. Chicago, IL: The University of Chicago Press.

Traugott, Elizabeth C. 2008. "All that he endeavoured to prove was …": On the emergence of grammatical constructions in dialogual and dialogic contexts. Pre-publication version of paper published in Robin Cooper & Ruth Kempson (eds.) *Language in Flux: Dialogue Coordination, Language Variation, Change and Evolution*, 143-177. London: Kings College Publications.

Weinert, Regina & Jim Miller. 1996. Cleft constructions in spoken language. *Journal of Pragmatics*, 25. 173–206.

Yatsiv-Malibert, Ilil. 2009. *le'ifyunam shel mivnim mevuka'im basafa hadvura* ['On cleft constructions in the spoken language']. *Balshanut 'ivrit* ['Hebrew Linguistics'] 62–63. 131–143.

Appendix: Transcription and glossing

Transliteration follows conventional *IPA* values, except for *y* for /j/, sh for /ʃ/, and an uninverted quotation mark (') for the glottal stop phoneme.

The transcription conventions of the Haifa Multimodal Corpus of Spoken Hebrew are based on those of the Santa Barbara Corpus of Spoken American English (Du Bois et al. 1992, Du Bois unpublished manuscript 2012), as adapted for Hebrew (Maschler 2017). Each numbered paragraph denotes a single intonation unit (Chafe 1994):

. . – perceptible pause of less than 0.1 second
. . . – average pause ($0.1 \leq x < 1.0$ s)
. . . . – pause ($1.0 \leq x < 1.5$ s)
. . . . . – pause ($1.5 \leq x < 2.0$ s)
(3.56) – measured pause of 3.56 s
, – comma at end of line – mid-level, mid-rise, mid-fall intonation, regularly understood in Hebrew as "more to come"
. – period at end of line – low fall intonation, regularly understood in Hebrew as final
? – question mark at end of line – high rising intonation, regularly understood in Hebrew as final and seeking response from interlocutor ('appeal')
?, – question mark followed by comma – rising intonation, regularly understood in Hebrew as projecting "more to come" while seeking (minimal) response from interlocutor
! – exclamation mark at end of line – final exclamatory intonation
∅ – lack of punctuation at end of line – a fragmentary intonation unit, one which never reached completion.
-- – two hyphens – elongation of preceding sound
- – one hyphen – morpheme boundary
<u>un</u>derlined syllable – primary stress of intonation unit
boldfaced syllable – secondary stress of intonation unit
@ – a burst of laughter (each additional @ symbol denotes an additional burst)
[Square bracket to the left of two consecutive lines indicates beginning of overlapping speech, two speakers talking at once

Alignment such that the right end of the top line
is placed over the left end of the bottom line indicates latching, no interturn pause.

Multimodal aspects of interaction (in green font) are rendered using the conventions developed by Mondada (2019). Specifically:

`+nod` A punctual embodied action is described following a single symbol in the line following the transcription (one symbol per participant and per type of conduct), synchronized with an identical symbol in the corresponding stretch of talk.
`§pointing§` Prolonged embodied actions are described between two identical symbols.
`---->` Described embodied conduct continues across subsequent lines
`----§` until the same symbol is reached.
`>>---` Described embodied conduct begins before the line's beginning.
`--->>` Described embodied conduct continues until the end of the excerpt.
* Exact position in the utterance in which a video caption was made.

CONTACT
maschler@research.haifa.ac.il
hillapolak@gmail.com

Linguistic Typology at the Crossroads 5-2 (2025): 161-187 ISSN 2785-0943

Sociolinguistics meets typology: Insight from vernacular speech to account for cross-linguistic patterns

SALI TAGLIAMONTE
UNIVERSITY OF TORONTO

Submitted: 05/06/2024 Revised version: 02/08/2025
Accepted: 07/08/2025 Published: 25/02/2026

Abstract

This study explores how two English grammatical systems with alternative grammatical options expose typological tendencies. A study of negation, *no* vs. negative quantifiers (*nothing*) vs. negative polarity items (*any/anybody*) reveals that the syntactic patterning of English indefinites and negative objects lines up with the hard syntax of the same expressions in closely related languages. A study of variable possession (*the* vs. *my/our*) reveals that while much of the system is stable, but nuances of contrast between UK and Canadian patterns expose differences in conceptions of the personal domain that relate to the cultures in which the varieties are spoken. Taken together these results make several important contributions: 1) the value of natural speech data in providing insights into linguistic and sociolinguistic typology operating in tandem; 2) the importance of statistical tools to uncover relevant patterns and contrasts in variable data. The findings affirm that patterns in single variety, when viewed across communities, dialects and languages reveal that linguistic systems are a continuum of interlocking structural and tractile contrasts in the broader context of human language.

Keywords: sociolinguistics, typology, negative quantifier, negative polarity item, possessive 'the', possessive pronoun.

1. Introduction

In this paper, I explore how vernacular speech data exposes cross-dialectal contrasts, which can, in turn, be used for insights into typology and vice versa. Comparative sociolinguistics (Tagliamonte 2002) and variationist analysis (Tagliamonte 2006;

2025) provide the theoretical approach and methodological toolkit, offering a quantitative, empirical means to understand the systems underlying language variation.

Using two different grammatical systems as case studies, I demonstrate how cross-dialectal regularities expose typological tendencies. A study of negation, (1), reveals that the soft syntax of English (eng) indefinites and negative objects aligns with the hard syntax of the same expressions in Scandinavian languages with object raising (Burnett et al. 2018: 103). (2) a study of possession of prototypical possessums (Gardner & Tagliamonte 2020) reveals a pattern found in other Indo European languages whereby nouns that are inherently possessed ('legs' and 'arms') do not require overt possessive marking (i.e., *my/our*) but can be marked with a definite pronoun, i.e., in English with *the.* Further, the analysis exposes a typological nuance between Canadian and British patterns based on varying cultural patterns of communal possession.[1]

(1)
a. There were ***no*** jobs to be had.
b. The were***n't any*** great places to eat.

(2)
a. 'Cause ***the*** house, I mean ***our*** house, it's not that big.
b. Had a picnic and went out on ***the*** bike.

1.1. The data

The data come from several sources: the Ontario Dialects Project (ODP), a 13-million-word archive of conversations with individuals from 21 communities across the largest province of Canada, Ontario. This is a long-term documentation project that has been underway since the early 2000's (Tagliamonte 2003–2006 et seq.).
Figure 1 shows a map of the communities in Ontario.

[1] All the examples come from the spoken language corpora under investigation. Note the local terminology and other linguistic features of interest in the examples, e.g., subject dislocation in (3).

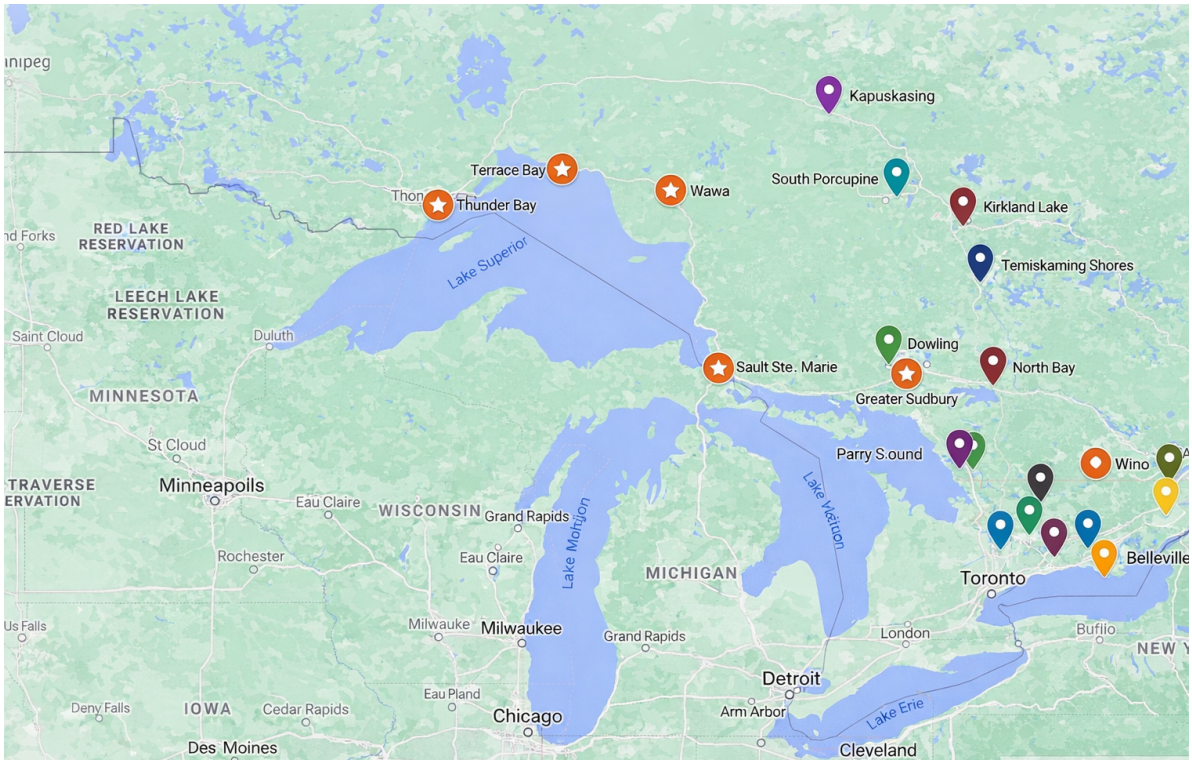


Figure 1: Map of Ontario showing the Ontario Dialects Project communities (Google maps).

In the UK, the data come from several different sources. The first, is the York English Corpus (YEC), a 1.2-million word archive of conversations with individuals born and raised in York collected in 1997 (Tagliamonte 1996–1998). The other corpora include a 3 million word archive of UK dialects comprising elderly people in small communities across the UK (ROOTS), (Tagliamonte 2001–2003; 2013) and a multitude of small studies conducted by students at the University of York from communities across the UK between 1997-2003 (Tagliamonte 2000–2001). In collaborative research, these materials have been supplemented from corpora collected in other universities, in this work the NECTE corpus, a compendium of vernacular speech collected in Newcastle, UK (Allen et al. 2007; Beal et al. 2007). Figure 2 shows a map of the communities in the UK.

All the materials in the synchronic corpora were collected through on the ground fieldwork and lengthy informal recordings of conversations with people born and raised in the communities. Due to legacy recordings in some locales, the time depth of these materials is substantive, with individuals born from late 1800's (from legacy recordings) to individuals born in the early 2001's. Further, because the corpora comprise geographic, social, and linguistic contrasts, the analyst can probe language features in these materials from varying perspectives. In this paper, I focus in on the

value of these materials for typology highlighting findings that have broad relevance; however, the scope of the data has to potential to explore many types of contrast.

Figure 2: Map of the United Kingdom showing locations of fieldwork (Google maps).

1.1.1. Can sociolinguistic data be useful for the study of typology?

Sociolinguistic data is spoken interaction in informal settings. It is often not ideally constituted, regular across places or data sets, or equally 'good' on many dimensions. When there is more than one person in the interview, they often talk over each other and the sound quality of a recording can be hampered by unexpected local conditions, e.g., grandfather clocks, aquariums, dogs, etc. The data are not balanced, as in corpus linguistics where there is a data extraction plan and cut-off point so that the number of words by category (e.g., register) is equal. This type of methodology is not the case in sociolinguistic data. The fieldworker aims for approximately an hour of conversation with each community member; however, it may be less or more. Another attribute of sociolinguistic data is that it is unpredictable. The fieldworker cannot anticipate what someone will talk about or for how long or what extraordinary words

and expression may come out of their mouth, or not. Finally, sociolinguistic data is crosscut by innumerable regional, social, and cultural factors.

Nevertheless, the Canadian and UK repositories are among the largest corpora of vernacular speech in the world and cover at least 120 years in birth dates of the individuals from the late 1800's up to the early 2000's. The data are sufficient for finding and studying linguistic variables that are notoriously rare in spoken corpora and may never appear in written sources. Comparative studies offer the ultimate check on interpretation and for discovering nuances both linguistically but also historically and culturally. Moreover, these materials offer a comparative perspective that is a must when considering issues of typology.

1.2. The value of social information and linguistic theory

I will argue that sociolinguistic metadata such as date of birth, gender, education, and job type have the potential to offer important insights into typology. Further, the different types of communities across Ontario and the UK afford the ideal opportunity to contrast varieties and communities where local features may expose unique forms and patterns. While a sociolinguistic approach to typology is not new, its application has mostly been applied to historical texts which have a strong bias towards the written language, formal settings and literate individuals (Walkden et al. 2023: 550). Most corpus-based studies come from sources other than the spoken language, including dialect surveys, atlases etc. Even when spoken language is investigated the corpora under investigation often do not provide representative coverage of community level social contrasts nor metadata about the characteristics of the individuals (e.g., Szmrecsanyi & Grafmiller 2023).

Building on the premise of the journal "Linguistic Typology at the Crossroads" whereby cross-linguistic tendencies are typically found in discourse use as pathways of language change (Keller 1994; Hawkins 2004; 2014; Haspelmath 2021), I suggest that social information, e.g. gender, occupation, education alongside geographic and regional differences, e.g. population size, distance from urban centers, can add depth and insights to these trajectories.

I also build on theoretical linguistic research that has demonstrated grammatical contrasts that are portrayed as 'hard' vs. 'soft'. Hard contrasts are the structural properties that create grammaticality differences between languages, the mainstay of typology, whereas soft contrasts are preferential tendencies between languages (e.g.

Bresnan et al. 2001). For example, in syntax languages with more flexible word order may allow for different types of alternations compared to languages with rigid word order. Similarly, in semantics, the meanings of individual words and how they interact within the meaning system of a language in conjunction with the cultural ecology in which they are used can also influence the types of alternations that occur. The same is true of phonological, morphological, and discourse-pragmatic patterns. I will argue that nuances in the constraints that operate on variable systems provide an ideal mirror of these soft contrasts and offer evidence for typological patterns. Further, trajectories of change can be mapped to geography (e.g. Poplack & Tagliamonte 1999) and sociolinguistic typology (Trudgill 2011; Szmrecsanyi & Wälchli 2014; Tagliamonte & Brook 2016; Torres Cacoullos & Travis 2019), although depending on the variable, these may or may not apply.

1.3. The value of the spoken vernacular

The spoken vernacular, what has been referred to as "naturally occurring speech" or "what language users do", is argued to be the "gold standard" for typology (Du Bois & Troiani 2023). Moreover, Miestamo et al. (2023) suggest that by "tapping into the rich vein of functional diversity found in naturally occurring conversation, typology will be better positioned to discover new forms of structural diversity". In the same way, dialect data are ideally positioned to add to the knowledge base of typology. Because of their low status dialects are often regarded as degenerate versions of the prevailing linguistic norms, therefore, they tend to preserve linguistic phenomena that are not recorded in descriptive accounts and so many of their features remain unattested in the published literature. The linguistic phenomena that endure in dialects offer insight into earlier typological structures in the same language as well as their history and culture that led to their evolution (Du Bois & Troiani 2023).

The deep knowledge that lives within community members as they use their vernaculars emanates the "imprint of ages emblazoned in vocabulary and expressions" (Tagliamonte 2016). Indeed, the words and expressions people use provide information far greater than their conscious knowledge of their language. From the broader contemporary context (currently in the early decades of the 21st century), it is also important to emphasize that dialects in rural locales all over the world are dying rapidly. The type of sociolinguistic fieldwork and documentation of research programs focusing on regional dialects aims to bring the linguistic treasures

held within dialect data into the open and into academic study as an enduring legacy for future generations. For the purposes of typology, this information has the potential to offer linguistic details of form and structure that are not represented in more formally gathered language materials.

1.3.1. A gold mine of linguistic data

Spoken vernacular dialect data collected in community context (i.e., representing age, gender and other socially relevant information) provides a cornucopia of sociolinguistic contrasts. Moreover, due to the wide range of birthdates of the individuals who live in geographically diverse locations, the Canadian and UK archives are replete with remnants of receding and emerging features. Conservative features from earlier times include zero articles, (3a), double demonstratives, (3b), old preterit forms, (3c), etc. and discourse-pragmatic markers, (4). The most infamous discourse-pragmatic marker in contemporary English is *like* (4a) (e.g., D'Arcy 2017), but other markers are developing (4b) and there are many old ones, (4c-d).

(3)

a. My mother's father was ***Ø relatively prosperous*** *farmer*. (mschiff, woman, 78, Parry Sound)
b. One day, we had ***this here snowstorm*** down here. (dhinds, man, 77, Christie)
c. I ***come*** home to Tatlock and I stayed there and ***run*** the mail. Looked after the post office. (mcarson, woman, 81, Ottawa Valley)

(4)

a. I think ***like*** I don't know ***like*** I guess.
b. No ***wait***, we did. ***Wait***, did we? .
c. ***See***, it would start about eight o-clock in the evening and go 'til about four o-clock in the morning, ***you see.***
d. If you did get onto the roads, ***why***, you never knew when you were going to get home ***or anything like that.*** (gbillings, man, 78, Ottawa Valley, 0101199_5)

Global mobility has led to a myriad of new language contact situations that — at the turn of the 21st century —can transcend geography offering new uncharted territory

for investigation, e.g. booming mega trends (e.g., Tagliamonte & Smith 2021) and so called "Black Swans" (Taleb 2007; Tagliamonte et al. 2016), that defy historical predictions about change.

In summary, spoken language data, messy and unpredictable as it is, offer the potential for deep insights into typology by seeking out rare phenomena in the dialects of closely related languages and deviations within types across varieties and dialects.

2. Crosslinguistic variation

As introduced earlier, recent research in linguistic theory has discovered that grammatical contrasts that distinguish languages can present as preferential tendencies *within* languages, suggesting that there is a critical connection between the syntactic structure and semantic organization of complex expressions and how they are used by speakers. A phenomenon that is categorical in one language, may be optional or variable in another.

Setting the scene for this kind of an approach, Bresnan et al. (2001) report on a comparison of person hierarchy effects and grammatical voice in Lummi, a Salish language compared to English. In Lummi, transitive predicates with third-person actors and first-person patients must be in passive voice; however, first or second person actors and third person patients must be in the active (Jelinek & Demers 1983). I illustrate with English equivalents in (5).

(5)

a. I am known by the man
   1st person 3rd person

b. The man knows me
   3rd person 1st person

In English, a third person actor with a first person patient is grammatical with either active or passive voice; however, in a study of these alternates in spoken language in the Switchboard corpus (Godfrey et al. 1992), notable contrasts were discovered. First and second person actors with a third person patient occur only in the passive

(n = 0/6246 cases) and third person actors with first or second person patients occur overwhelmingly (97%) in the active (n = 472/486). Although both options are grammatical, an active and passive contrast is operational in usage patterns. The critical facet of this result is that it demonstrates that hard contrasts in some languages may manifest as soft contrasts in others, a finding that has been repeated for other features of English, such as the dative alternation (e.g. Bresnan et al. 2007; Bresnan & Ford 2010) and the animacy effects in the genitive alternation (e.g. Zaenen et al. 2004; Wolk, et al. 2013; Szmrecsanyi et al. 2017; Szmrecsanyi & Grafmiller 2023). If we turn these observations into a workable hypothesis then an important question to ask is: "to what extent do the apparently variable English patterns have correspondences with invariant syntactic patterns in other languages?" (Burnett et al. 2018: 91).

3. Two case studies of variation

In the next section, I highlight the results of two case studies with the goal of offering insights from large archives of sociolinguistic data to the study of typology. The two case studies are excerpted from research from the corpora described in section 1.1: 1) grammatical choices in negative patterns with negative quantifiers and polarity indefinites (Childs et al. 2015; Burnett et al. 2018) and 2) the alternation between definite article and possessive pronouns (Gardner & Tagliamonte 2020). In both case studies, there are two standard options that mean more or less the same thing, i.e. ***I don't know anything/I know nothing*** and ***my bike/the bike,*** thus constituting a 'linguistic variable' (Labov 1972 et seq.). Further details on situating these linguistic phenomena in time and space, details of methodology and results can be found in the research papers.

3.1. Negation patterns

In English negative constructions, negative quantifiers (e.g., *no, none, nothing, nobody,* etc.) and negative polarity indefinites (NPI's) (e.g., *any*, *ever, at all*) alternate in constructions where either one is grammatical, (6a-b).

(6)

a. *There were* ***no*** jobs to be had. (There *were****n't/not any*** jobs to be had)

b. *The were****n't any*** great places to eat. (There were ***no*** great places to eat)

These alternates occur with near equal frequency in naturally occurring speech. The question is what determines whether an individual will use one alternate vs. the other?

Many studies — most based on historical written materials — report a consistent pattern. A hierarchy of verb constructions conditioning the options: existential *be* occurs most often with negative quantifiers, followed by stative *have*, then copula *be* and finally lexical verbs, which tend to prefer negative polarity indefinites, also referred to as NPI's. Notice however, that both options are grammatical in each category, as in (7) with NPI's and in (8) with *no*.

(7) Existential *be* → stative *have* → Copula *be* → Lexical verbs

a. There *wasn't* ***anything*** else to do. (prowlett, man, b. 1943, Toronto, Ontario)[2]
b. They *haven't* acquired ***anything*** yet. (pronenen, woman, b. 1955, Toronto, Ontario)
c. She *was no* good for ***anybody***, that woman. (D/131, man, b. 1931, Newcastle, UK)
d. They *don't look* ***any*** different. (echapman, woman, b. 1942, York, UK)

(8) **Existential *be* → stative *have* → Copula *be* → Lexical verbs**

a. There *was* ***no*** comaradship. (pmarshall, man, b. 1949, Wheatley Hill, UK)
b. My tractor *had* ***no*** brakes, ***no*** nowt. (crabeale, man, b. 1967, York, UK)
c. I *'m* ***no*** different now. (rburkett, female, b. 1968, Belleville, Ontario).
d. My grandmother *spoke* ***no*** English. (oholtby, b. 1951, Toronto, Ontario)

In Childs et al. (2015), we examined the patterns of use of these alternates in individuals born and raised in two Canadian communities, a large city (Toronto) and a small town (Belleville) along with and several northeastern communities in the UK, York and Newcastle (two cities), Durham (a small city) and Wheatley Hill (a small town).

Figure 3 provides a more nuanced perspective, distinguishing Canada vs. the UK and showing the overall rates of negation in each community.

[2] Each example indicates the individual's pseudonym, their perceived gender, date of the birth and place of residence.

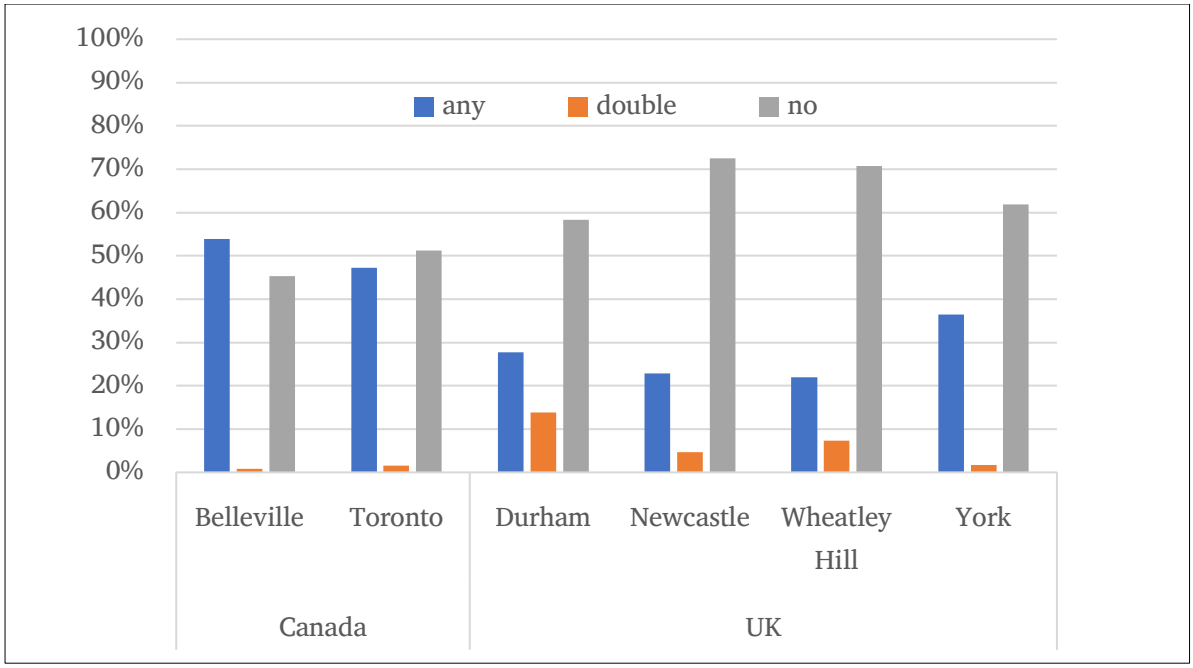


Figure 3: Rate of negative quantifiers across communities in Canada and the UK.

Double negation (i.e., negative concord) is virtually absent in Toronto and Belleville as well as in York. It also occurs rarely (6.6% of the time) in Newcastle, Wheatley Hill, and Durham. In contrast, variation between *no*- and *any*-negation is present in all varieties, but the distribution is markedly different by variety. In Canada, the two constructions have near-equal frequency, with a slight preference for *no*-negation in Toronto. In England, *no*-negation dominates at 63% in York and ranges from a high of 73% in Newcastle to 58% in Durham. Given that *no*-negation is the older form, it appears that UK dialects, at least in the northeast are more conservative than Canadian dialects in Ontario. However, there is notable inter-variety coherence in overall rates of variants by community regardless of population size or location.

In the next step, we replicated the coding schema of earlier studies and tested for the effect of the hierarchy of verb/contruction types that had been reported in studies of historical written English. Figure 4 shows the rate of *no*-negation according to these construction types, amalgamating the data in the UK as "Modern Spoken" and in Canada as 'Toronto, Canada' and comparing it with the rates reported in earlier studies of written data from the UK (Early Modern and Modern English).

The construction hierarchy in the UK and CDA data mirrors, in large part, the one found in Early Modern and Modern Written English. Each data set exhibits the same overall hierarchy: higher rates of negative quantifiers, i.e. *I know nothing*, with existential *be*, then copula *be*, *have* and lexical verbs the least, i.e., lexical verbs tend

to occur with *any* negation, i.e. *I don't **like** anything*. However, notice that rates of negative quantifiers are lower overall in spoken data.

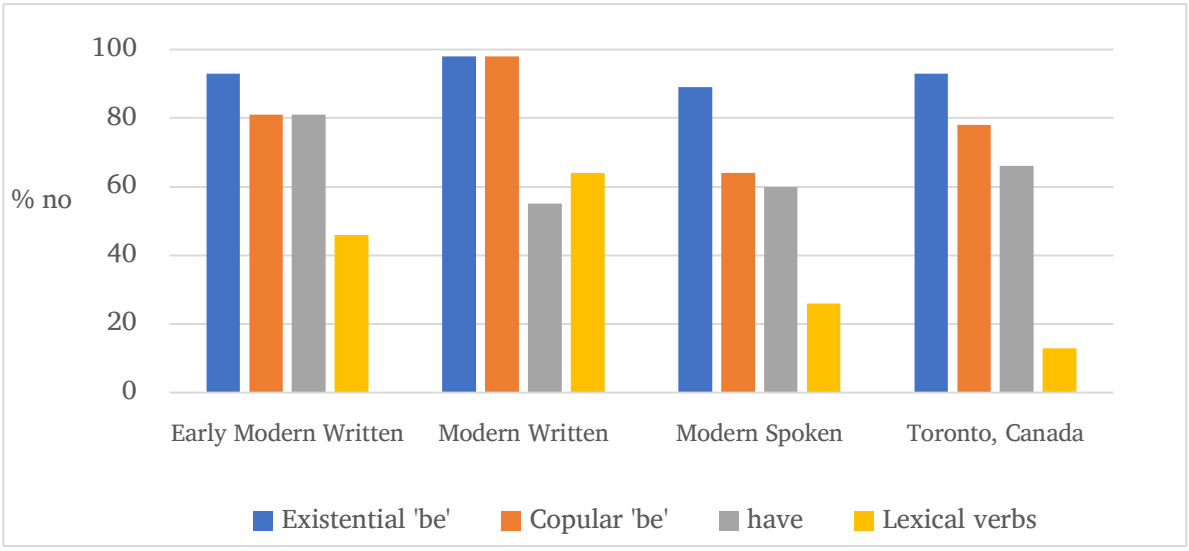


Figure 4: Rate of negative quantifiers in written compared to modern spoken English.

Moreover, Early Modern English differs from Modern English in having greater use of negative quantifiers with *have* and 2) the modern spoken data from Canada has far greater use of negative polarity items with lexical verbs. Note too that the spoken British and Canadian English data are much the same, except for the nuance that the UK data has higher use of negative quantifiers with lexical verbs, (9) than the Canadian data where it is rare.

(9) I *see **no*** point in it. (JS/169, man, b. 1982, Newcastle)

These differences mean that a frequency-based explanation does not entirely explain the variation and that deeper consideration of these differences was warranted.

In a later study focusing on the same negative options, we augmented the analysis by collaborating across disciplines, adding a socio-semanticist with a specialization in mathematical tools (Burnett) and a formal syntactician (Koopman) to the research team. Then we went back into the data to conduct a more in-depth analysis. The question we aimed to address was: *what other linguistic patterns may underlie the linear order of the constraint hierarchy?*

3.1.1. Expanding the perspective to typology

To re-focus attention on underlying structural patterns, we scrutinized languages related to English for insight, for example, Scandinavian languages. Kayne (1998; 2000) demonstrated the key syntactic contrast of object raising that distinguishes negative quantifiers and negative polarity items based on the nature of the verb. In English the syntactic position occupied by object negative quantifiers differs according to the syntactic properties of the other morphosyntactic material that the indefinite combines with. For example, in some verbal constructions, such as an existential (10a), the direct object *nothing* has undergone a negative quantifier shift to a higher syntactic position than it would occupy if it had occurred in a structure with a lexical verb (10b) or a participle (10c).

(10)

a. There'*s* *nothing* to do here. (katrina, woman, b. 1967, York, UK)
b. He eats *nothing*. (AA/613, man, b. 1985, Newcastle, UK)
c. She got her degree from Ryerson. She*'s* *doing* *nothing* with it. (akaran, woman, b. 1979, Toronto, Cda)

In other words, only the verb *be* and the object can raise out of a verb phrase. In contrast, other verbs cannot raise, so the object does not raise either, (11).

(11)

a. There are *no* → Neg-Q
   There are *no* absolute rules now. (ekempt, woman, b. 1928, Toronto, Ontario)

b. There *aren't* ***any*** … → NPI
   There *are****n't*** a lot of mosquitos. (asinkic, woman, b. 1962, Toronto, Ontario)

Based on Kayne's analysis, we predicted that when it is possible for the object to move upwards in the syntactic structure, negation is more likely to be realized with negative quantifiers.

Figure 5 shows a notable contrast: where the negative is higher than VP, a negative quantifier is used; where it is lower than VP it appears as negative polarity item. There is only a small amount of variation.

3.1.2. Results of typology on variation — negation

We coded the Ontario and UK data (see section 1.1) according to whether there was object raising above the verb phrase or not (see Burnett et al. 2018 for coding and other information). The results are illustrated in Figure 5.

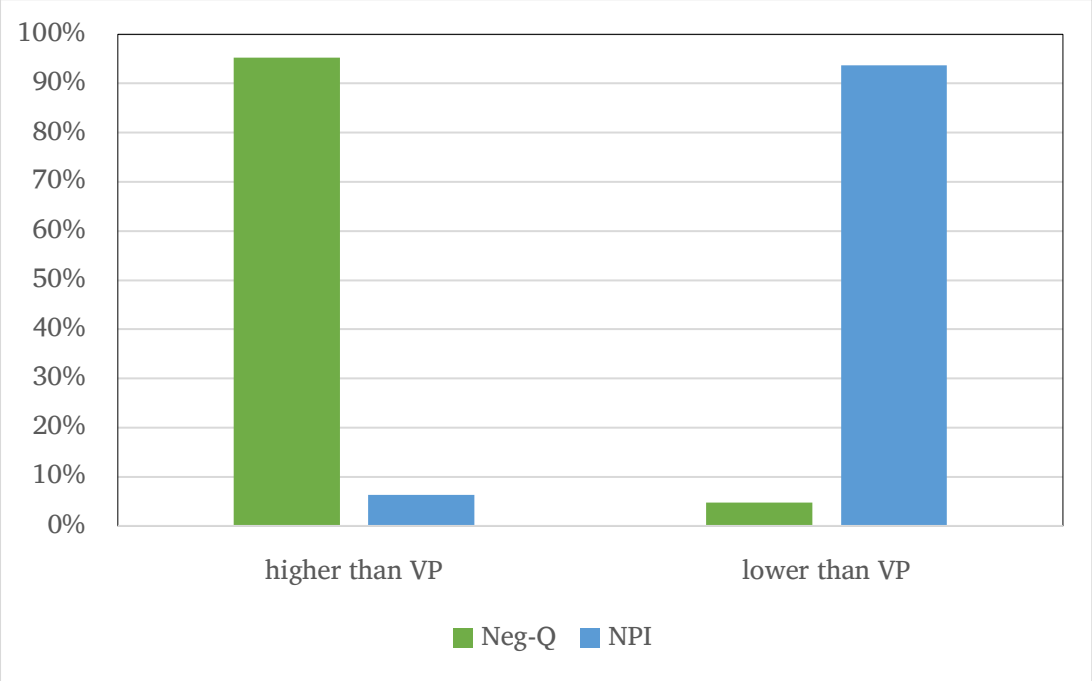


Figure 5: Rate of negation constructions by syntactic domain.

3.1.3. Summary

The key finding is that when the negation constructions are coded according to syntactic domain rather than construction type, the variants are nearly completely distributed according to syntactic position: *no/not* appears in the higher syntactic domain, 95% and negative quantifiers (e.g., anything) appear in the lower syntactic domain, 95%. The variation is highly circumscribed even though both options are grammatical in the language.

The cross-linguistic regularities from the comparative syntactic study show that the structural position that the negative quantifier or negative polarity it occupies plays and important role in negative patterns in English. A hard contrast found in some Scandinavian languages is manifested as a soft contrast in English.

In section 3.2. I turn to the second case study, personal domain possessums.

3.2. Personal domain possessums

English offers two possibilities for marking possession in the personal domain, a definite article, (12a-b) and a possessive determiner, (13a-b), and both options are grammatical.

(12)

a. ***The*** *house*, I mean ***our*** *house*; it's got- it's not that big. (mlondry, woman, b. 1959,York, UK)

b. I was at home and ***my kids*** couldn't eat the ice-dream fast enough out of ***my freezer***. … … ***Our*** cottage is like a cottage from years ago. … I mean we light fire and ***the kids*** still do their hot-dogs and stuff … (rsinardo, woman, b. 1964, Toronto, Cda)

Where does an English speaker use *the* and when *my/our?* Upon reflection, even native speakers, wonder where one or the other form should be used; yet in unreflectingly vernacular conversation both options occur regularly. Grannis (1972) lamented that "the major modern theoretical approaches to language description [have] not yet developed [the] means of adequately accounting for the referential complexities involved in the use of the definite article." He called this the "Definite Article Conspiracy". To illustrate, Grannis (1972) elaborated that a man with a missing cat might say to his wife, "have you seen *the* cat?" If she hasn't, he might then ask his neighbour, "Have you seen *our* cat?" To a complete stranger, he might then say, "Have you see *a* cat?" But to any one of these people the same man could say, "I have to feed *the* cat." Likewise, to your wife, you could say, instead of "*the* cat", "where's *our* cat?" Furthermore, you could also call your wife or husband, *"the* wife/*the* husband*"*. This description underscores the robust variability of forms.

In Gardner and Tagliamonte (2020), we studied the patterning of these options in a comparative study of spoken vernacular discourse in the UK and Canada using the York English Corpus and the Toronto English Archive (section 1.1). We based our data extraction using the concept of personal domain (Bally 1926) in which membership in the personal sphere of the speaker is the relevant dimension of contrast (McWhorter 2002: 224). We focussed on nouns "conventionally understood to be within the personal domain" (Bally 1926: 233). Common or prototypical words from several personal domain-type semantic categories (e.g. body parts, intimates,

conveyances, etc.) were used as search keywords, including *house, back garden/yard, car, boyfriend/girlfriend,* etc.), (13a-b).

(13)

a. The only thing I've ever gotten stolen was – when I– the day I got married, someone stole a case of beer from ***our*** *backyard*, someone stole gas out of ***the*** *car*. Siphoned it out of ***the*** *car*. (njalna, woman, b. 1964, Toronto, Cda)
b. We had to take ***the*** car for some work doing and we use a chap in Elvington and … so I drove ***our*** car and Wayne went on his motorbike. (cbiggs, woman, b. 1964, York, UK)

Dictionaries and usage guides consistently report that when possessive *the* is used the context is informal, (14a-c).

(14)

a. (*informal*) used instead of a possessive pronoun to refer to someone with whom the speaker or person addressed is associated: *I'm meeting the boss | How's the family?* (Oxford Dictionary of English, 2nd edition revised)
b. British informal *my, our* (Oxford Dictionary of Canadian English)
c. Colloquial: *my:our, the dog, the car* (Oxford Dictionary of Current English)

Possessive *the* is also said to be used only to refer to someone who the speaker or interlocutor is associated with such as "*the* boss" or "*the* family". It may also used for body parts, (15a-b). Notice here that the possessum — the thing possessed — is restricted to possessions of humans.

(15)

a. *how's the arm today?* used as a function word before names of some parts of the body or of the clothing as an equivalent of a possessive adjective (Merriam-Webster)
b. used to refer to a part of the body, e.g. *Lieutenant Taylor was wounded in* ***the*** *knee; How's* ***the*** *ankle? Is it still hurting?* (Longman Dictionary of Contemporary English, Advanced Learner's Dictionary)

Possessive *the* is also said to be related to the habit of men referring to their wives, or

children with *the*, as in "How's *the* wife" or "How are *the* kids?" It is also a familiar way of referring to a husband or father, as in "*the* old man". However, such use is restricted to "some men", perhaps suggesting social stratification as well as stylistic correlates of "informality" and "familiarity". It may also be a regional dialect difference.

In summary, both linguistic and social factors are reported to be involved in the usage patterns of possessive *the* vs. posessive pronouns *my/our* in English with personal domain possessums.

3.2.1. Expanding the perspective to typology

To focus on underlying structural patterns, we scrutinized languages related to English for insight. In this case the relevant comparison is the Indo-European language family where a special relationship between definiteness and possession is documented. First, possessed nominals that are marked with a possessive pronoun do not require overt definite marking because they are inherently definite. Second, objects that are considered "inherently possessed" — like body parts — do not need overt possessive marking and usually only appear with definite marking. This contrast is a hard contrast in several languages like French (fra) and Italian (ita), as in (16a-b). However, across languages the patterns differ slightly and in languages such as Spanish and English, there is alternation.

(16)

a. French
Je me lave ***les*** *mains*
I to.myself wash ***the*** *hands*
'I wash my hands'

b. Italian
mi lavo ***le*** *mani*
to.myself wash ***the*** hands
'I wash my hands'

Our goal, therefore, was to uncover the patterns underlying the choice between *the* vs. *my/our* in Canada and the UK as represented by Toronto and York in a set of objects within the personal domain across semantic categories.

Because alienability is basically the distinction between body parts and everything else, we adopted Bally's concept of "personal domain" in order to extend the coverage to a more fine-grained categorization schema.

Bally considered personal domain to be entirely subjective:

> nothing prevents the collective imagination from attributing to the self objects that normally have their own independent existence, or, conversely, of detaching those things which in reality cannot be. The extent of the domain is determined by the cultural outlook of each linguistic group. Its limits may vary from language to language and vary within the same language during the course of its evolution.

Therefore, we left open the possibility that the choice may not only be linguistic but also determined by the varietal differences and cultural context in time and space.

3.2.2. Results of typology on variation — possession

First, we discovered that the rates of possessive *the* and possessive pronouns *my/our* were remarkably similar in Toronto and York. Overall possessive *the* was used in about 30% of the possessive personal domain contexts we included in the study. Then we performed a mixed-effects logistic regression model in R (R 2022) to test the statistical validity of the many potential social and linguistic patterns that had been mentioned in the literature, while treating individual as random.

We discovered that most of the predictors that had been attested in the literature were not significant. Grammatical person was not significant, nor was syntactic position of the possessum, its grammatical number, whether it was human or not, animate or not or alienable or not. Notably, alienable possessums, e.g. *dogs, kayaks,* were more likely to be used with possessive *the* than inalienable possessums, e.g. *arms, bellies,* despite the claim in Quirk et al. (1985) that possessive *the* is idiomatically preferable to a possessive pronoun in inalienable contexts.

Instead, alienable possessumes had a higher rate of possessive *the* than inalienable possessums, the reverse of what was predicted. With respect to the social factors, men used slightly more possessive *the* in Toronto and women use slightly more possessive *the* in York, but these differences were not significant. We found no trend in apparent time: younger speakers did not use less possessive *the*. Comparing speaker education and occupation did not reveal any significant patterning either.

The best fit of the model to our data categorized the possessums into semantically related objects, rather than an animacy contrast between human vs non-humans, or animates vs non-animates. Table 2 shows the rate of possessive *the* in Toronto and York across possessum categories in the study.

	Bikes	**Body parts**	**Family**	**Homes**	**Pets**	**Romantic partners**	**Vehicles**	**Yards**
				Toronto				
the	12%	9.9%	26.1%	51.1%	75%	4.5%	57.2%	49.7%
my/our	88%	90.1%	73.9%	48.9%	25%	95.5%	42.8%	50.3%
				York				
the	28.8%	5.4%	40.5%	47.3%	85.5%	3.92%	63.8%	76.58
my/our	74.4%	71.2%	94.6%	59.5%	52.7%	14.5%	96.1%	23.4%

Table 2: Rate of possessive *the* and possessive pronouns *my/our* by semantic category.

In both places, pets, and vehicles have the highest rates of possessive *the* and bikes, body parts. Homes, (e.g. cottages, cabins, and chalets) are robustly variable with both forms in both. Non-romantic family members (e.g. children, kids, and in-laws) are positioned in the middle between these groups, but note the striking difference in use of possessive *the* — Toronto (26.1%) and York (40.5%).

Another difference between York and Toronto is with 'yards': this category has one of the highest rates of possessive *the* in York (76.6%), but in Toronto yards have 49.7% possessive *the*, patterning with vehicles and homes.

3.2.3. Summary

The key finding is that there is remarkable similarity between Toronto and York in the rate and conditioning of marking possession with *the* vs. *my/our*. This parallelism extends to the lack of significant linguistic and social conditoning on the variation (see Gardner & Tagliamonte 2020: 246, Table 6).

Taken together these results suggests that the choice of possessive *the* and possessive *pronoun my/our* is a stable, soft contrast in English. However, rather than marking a hard contrast based in inalienability as in some Indo-European languages, in English it encodes communal possession. For example, pets, which occur with possessive *the* 75% and 85% of the time in this cross-variety comparison, are generally co-owned by

a group of people, such as a family in both cultures.

A similar personal domain interpretation can be expected for the family car or the family home. Body parts on the other hand only belong to one person and these are overwhelmingly encoded with *my/our*. Likewise, romantic partners like husbands and girlfriends are the least communally owned, and the ownership is more reciprocal, at least in contemporary western culture.

However, the interesting nuance to this linguistic system is that in two areas of personal sphere, the UK and Canada diverge considerably. In York 'yard' and other green spaces are very often referred to with possessive *the*. In Toronto, non-romantic family members (e.g. 'children', 'sister') are referred to overwhelmingly with possessive *my/our*.

4. Speculating on the impact of culture

The difference between 'cars' and 'bikes' is probably the clearest illustration of the striking regularity across dialects. Figure 6 shows the rate of possessive *the* in 'cars' vs. 'bikes' in Toronto (Canada) and York (United Kingdom).

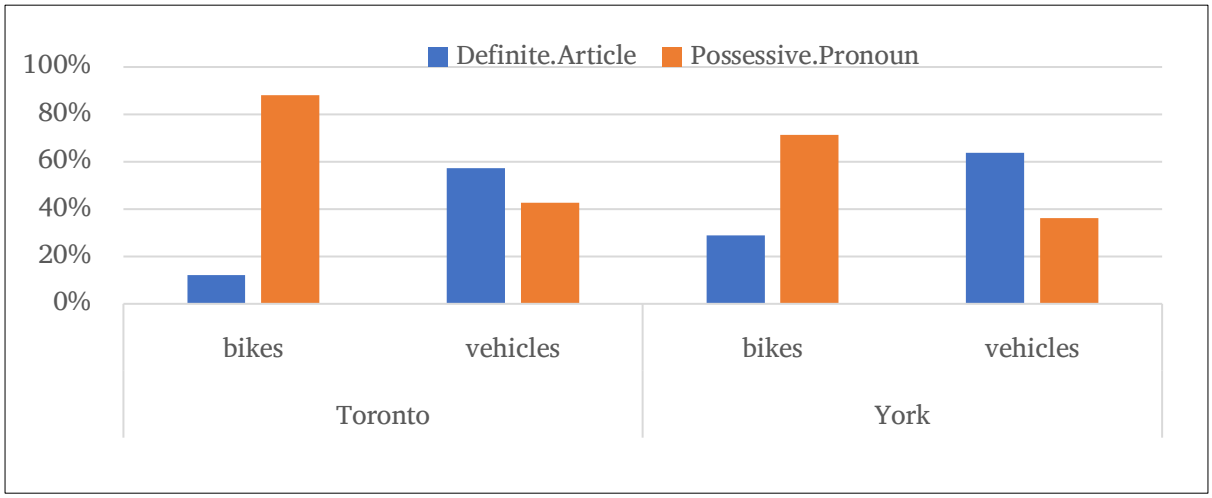


Figure 6: Rate of possessive *the* vs. possessive pronouns *my/our* with bikes vs. vehicles.

Cars (vehicles) which are generally thought of as belonging to a family occur with possessive *the* about 50% of the time, while bikes, which generally belong to only one person, have more *my* and the contrasts across varieties is parallel. In contrast, Figure 7 also highlights a striking deviation between the two dialects, namely the rate of possessive *the* with yards and children in Toronto compared to York.

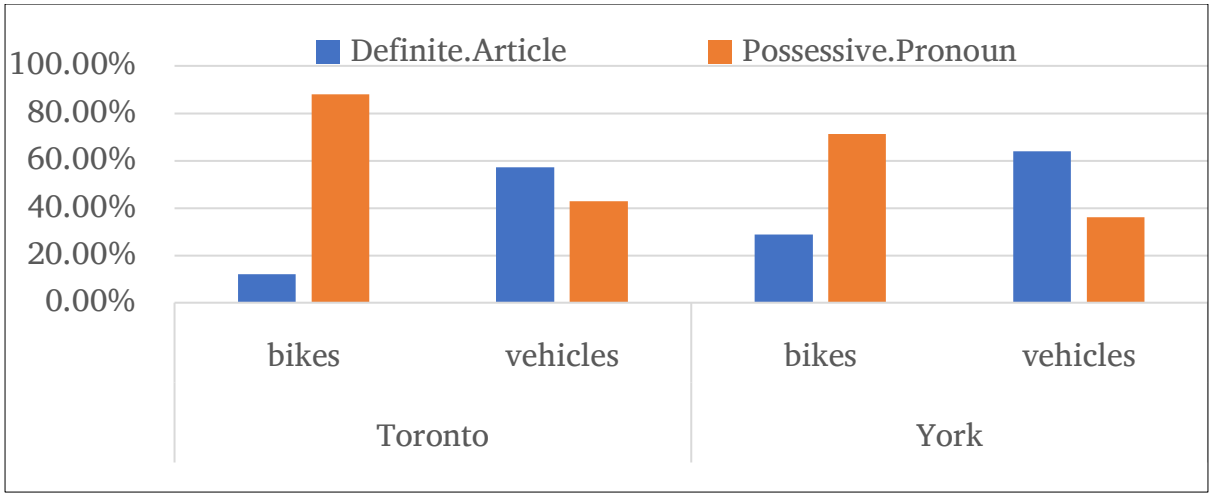


Figure 7: Rate of definite article *the* vs. possessive pronouns *my/our* with bikes vs. vehicles.

The patterns here are antithetic. In Gardner & Tagliamonte (2020:249), we speculated that the higher rate of possessive *the* for yards (76% compared to 54%) among York speakers is likely due to the fact that the green spaces in York are spaces like the garden and the family allotment are spaces that are conceived as being more communal than the often expansive front and back yards of Torontonians. It is not clear however, why children, i.e. *the children* not *my children,* would by the use of *the* be construed as conceptually more shared in York than in Toronto is beyond the scope of the study, but is also suggestive of cultural differences. We can conclude that where cultures treat items in the same way, we can expect similar even universal patterns in the constraints on alternating options (i.e. *the* vs. *my/the*), but when cultures differ in their conception of items in terms of the personal sphere, we can expect differences.

This discovery aligns with earlier research in demonstrating that culture is a key attribute of typological variance (e.g., Evans 2010; Song 2010) and highlights a novel direction for study. Possession is subjective, varying across cultures and time (Bally 1926). These results suggest this malleability is susceptible to the ways space is conceived as well as in how human relationships are construed. If so, then there is also the possibility that the conception of personal domain is reflected in linguistic variation as one semantic category or another can shift over time and that these can be culturally constructed and locally distinctive. I suggest therefore that targeting linguistics variation in systems such as the marking of possession that are implicated in changing cultural norms and practises are ripe for diachronic study and will provide interesting fodder for future research.

The key finding is that when possessive constructions are coded according to personal domain, there was no correlation between the use of possessive *the* and the traditional users of non-standard speech, e.g., men or working-class speakers, as predicted in much of the related literature on this topic. Instead, we discovered that possessive *the* is typically used instead of a possessive pronoun (i.e., *my/our)* with possessums that are possessed communally (i.e., *the dog, the car*) rather than individually (*my bike, my boyfriend, my back*). In this case study, cross-linguistic comparison exposes a typological contrast in the personal domain, with nuances that reflect differences between two regional varieties. A striking finding is that there seems to be a critical link with culture in the variable systems of certain categories that warrant further investigation.

5. Discussion

Examination of two variable linguistic systems in spoken English language data have revealed insight into typology, in particular the relevance of hard and soft contrasts. In both linguistic systems, cross-linguistic regularities common in other languages were profitably used to understand why there is a choice, in one case between negative quantifiers and negative polarity items and in the other between possessive *the* and possessive pronouns (*my/our)*. In the former case, we identified a structural contrast that underlies an often-reported construction hierarchy in related languages. In the latter case, the marking of possession found a largely stable system. However, we also uncovered a dialect difference between the UK and Canada that exposed an area of 'softness' in conceptions of personal domain that are different. Importantly, what we know about variable systems is enhanced by expanding the knowledge base to include a typological perspective thereby increasing understanding and explanation more generally.

The two case studies also highlight the value of natural spoken language interactions collected using sociolinguistic methods to provide insights into typology. They also show that the tools of quantitative analysis and statistics applied to corpus data keyed to time and place, culture and context offer an important means to detect the details of typological patterns and how these play out in dialectal and cross-linguistic variation, all leading to more integrated explanations. Patterns in single variety, when viewed across communities and across dialects and languages confirm

that linguistic systems are a continuum of interlocking patterns in the broader context of human language.

Acknowledgements

I gratefully acknowledge the support of the funding agencies that enabled the research to be conducted, in Canada spanning 2001-present from the Social Science Research Council (SSHRC), in the UK spanning 1995-2003 from the Economic and Social Science Research Council (ESRC). I express my sincere gratitude to the individuals in the communities in the UK and Canada who shared their stories and reminiscences with me and my research team. The corpora were constructed by a skilled team of research assistants in the sociolinguistic laboratories in York and Toronto. Data extraction, coding and analysis for the two studies was made possible by collaboration with *Heather Burnett, Karen Corrigan, Claire Childs, Christopher Harvey, Hilda Koopman,* and *Matt Hunt Gardner*.

Abbreviations

CDA = Canada
NECTE = Newcastle Corpus of Tyneside English
NPI = Negative Polarity Item
YEC = York English Corpus
ODP = Ontario Dialects Project
UK = United Kingdom

References:

Allen, William & Joan Beal & Karen Corrigan & Warren Maguire & Hermann Moisl. 2007. A linguistic 'time-capsule': The Newcastle Electronic Corpus of Tyneside English. *Creating and Digitizing Language Corpora: Diachronic Databases*, 16–48. Basingstoke: Palgrave Macmillan.

Bally, Charles. 1926. L'expression des idées de sphère personnelle et de solidarité dans les langues indoeuropéenes. In Franz Frankhauser & Jakob Jud (eds.), *Festschrift Louis Gauchat*, 68–78. Aarau: Sauerländer.

Beal, Joan & Karen Corrigan & Hermann Moisl (eds.) 2007. *Using Unconventional Digital Language Corpora: Volume 1: Synchronic Corpora*. Basingstoke, Hampshire: Palgrave Macmillan Limited.

Bresnan, Joan & Shipra Dingare & Chris D. Manning 2001. Soft constraints mirror hard constraints: Voice and person in English and Lummi. In M. Butt & T. Holloway King (eds.), *Proceedings of the LFG01 Conference*, 13–32. Stanford: CSLI Publications.

Bresnan, Joan & Anna Cueni & Tatiana Nikitina & Harald R. Baayen. 2007. Predicting the dative alternation. In Gerlof Boume & Irene Krämer & Joost Zwarts (eds.), *Cognitive foundations of interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Arts and Sciences.

Bresnan, Joan & Marilyn Ford 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86(1). 168–213.

Burnett, Heather & Sali A. Tagliamonte & Hilda Koopman. 2018. Soft Syntax and the Evolution of Negative and Polarity Indefinites in the History of English. *Language Variation and Change* 30(1). 83–107.

Childs, Claire & Christopher Harvey & Karen Corrigan & Sali A. Tagliamonte. 2015. Comparative sociolinguistics insights in the evolution of negation. *Selected Papers from New Ways of Analyzing Variation (NWAV) 43*, Philadelphia, PA: University of Pennsylvania Working Papers in Linguistics, 21, Article 4.

D'Arcy, Alexandra. 2017. *Discourse-Pragmatic Variation in Context: 800 years of like*. Amsterdam and New York: John Benjamins.

Du Bois, John W. & Giorgia Troiani. 2023. *Typology and its data: Functional monoculture or structural diversity?* Naturally occurring data in and beyond linguistic typology. Bologna. 18–19 May 2023, Bologna.

Evans, Nicholas. 2010. 504 Semantic Typology. In Jae Jung Song (ed.), *The Oxford Handbook of Linguistic Typology*, 0. Oxford University Press.

Gardner, Matt Hunt & Sali A. Tagliamonte. 2020. The bike, the back, and the boyfriend: Confronting the “definite article conspiracy” in Canadian and British English. *English World-Wide* 41(2). 226–255.

Godfrey, John J. & Edward C. Holliman & Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. *Acoustics, Speech, and Signal Processing*, 517–520. New York: Institute of Electrical and Electronics Engineers.

Grannis, Oliver C. 1972. The Definite Article Conspiracy in English”. *Language Learning* 22. 275–289.

Haspelmath, Martin. 2021. Explaining grammatical coding asymmetries: Formfrequency correspondences and predictability. *Journal of Linguistics* 57(3). 605–633.

Hawkins, John. 2004. *Efficiency and complexity in gramamrs*. Oxford: Oxford University Press.

Hawkins, John A. 2014. *Cross-linguistic variation and efficiency*. New York: Oxford University Press.

Jelinek, Eloise & Richard A. Demers. 1983. The agent hierarchy and voice in some Coast Salish languages. *International Journal of American Linguistics* 49. 167–185.

Kayne, Richard S. 1998. Overt vs. covert movement. *Syntax* 1. 128–191.

Kayne, Richard S. 2000. *Parameters and Universals*. Oxford: Oxford University Press.

Keller, Rudi. 1994. *On language change: The invisible hand in language*. London and New York: Routledge.

Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

McWhorter, John. 2002. What happened to English? *Diachronica* 19(2). 217–272.

Miestamo, Matti & Ksenia Shagal & Olli Silvennoinen & Chingduang Yurayong. 2023. *Typology and usage: Explanatory perspectives with special reference to negation*. Naturally occurring data in and beyond linguistic typology. Bologna. 18–19 May 2023, Bologna.

Poplack, Shana & Sali A. Tagliamonte. 1999. The grammaticalization of *going to* in (African American) English. *Language Variation and Change* 11(3). 315–342.

Quirk, Randolph, & Sidney Greenbaum & Geoffry Leech & Jan Svartvik. 1985. *A comprehensive grammar of the English language*. New York: Longman.

Song, Jae Jung. 2010. Setting the Stage. In Jae Jung Song (ed.), *The Oxford Handbook of Linguistic Typology*, 0. Oxford University Press.

Szmrecsanyi, Benedikt & Jason Grafmiller. 2023. *Comparative variation analysis: syntactic variation in World Englishes*. Cambridge: Cambridge University Press.

Szmrecsanyi, Benedikt, & Jason Grafmiller & Joan Bresnan & Anette Rosenbach & Sali A. Tagliamonte & Simon Todd. 2017. Spoken syntax in a comparative perspective: the dative and genitive alternation in varieties of English. *Glossa* 86. 1–27.

Szmrecsanyi, Benedikt & Bernhard Wälchli (eds.) 2014. *Aggregating dialectology, typology, and register analysis: Linguistic variation in text and speech*. Berlin: de Gruyter.

Tagliamonte, Sali A. 2000–2001. Vernacular Roots: A database of British dialects. Research Grant B/RG/AN 6093/APN11081. Arts and Humanities Research Board of the United Kingdom (AHRB).

Tagliamonte, Sali A. 2001–2003. Back to the roots: The legacy of British dialects. Research Grant. Economic and Social Research Council of the United Kingdom (ESRC). #R000239097.

Tagliamonte, Sali A. 2002. Comparative sociolinguistics. In J.K. Chambers, Peter Trudgill & Natalie Schilling-Estes (eds.), *Handbook of language variation and change*, 729–763. Malden and Oxford: Blackwell Publishers.

Tagliamonte, Sali A. 2006. *Analysing sociolinguistic variation*. Cambridge: Cambridge University Press.

Tagliamonte, Sali A. 2013. *Roots of English: Exploring the history of dialects.* Cambridge: Cambridge University Press.

Tagliamonte, Sali A. 2016. *Making Waves: The History of Variationist Sociolinguistics*. Malden and New York: Wiley-Blackwell Publishers.

Tagliamonte, Sali A. 2025. *Analysing sociolinguistic variation, 2nd edition*. Cambridge: Cambridge University Press.

Tagliamonte, Sali A. & Marisa Brook. 2016. *Adaptive change in sociolinguistic typology: The case of relative 'who'*. LSA Annual Meeting 2016, January 7-10, 2016. Washington, D.C., USA.

Tagliamonte, Sali A., Alexandra D'Arcy & Celeste Rodrigues-Louro. 2016. Outliers, impact and rationalization in linguistic change. *Language* 92(4). 824–849.

Tagliamonte, Sali A & Jennifer Smith. 2021. Obviously undergoing change: Adverbs of evidentiality in the UK and Canada over 100 years. *Language Variation and Change* 33(1). 1–25.

Taleb, Nassim Nicholas. 2007. *The Black Swan: Impact of the highly improbable*. New York: Random House.

Torres Cacoullos, Rena & Catherine E. Travis. 2019. Variationist typology: Shared probabilistic constraints across (non-)null subject languages. *Linguistics* 57(3). 653–692.

Trudgill, Peter J. 2011. *Sociolinguistic typology: Social determinants of linguistic complexity*. Oxford: Oxford University Press.

Walkden, George & Gemma Hunter McCarley & Raquel Montero & Molly Rolf & Sarah Einhaus & Henri Kauhanen. 2023. Sociolinguistic Typology Meets Historical Corpus Linguistics. *Transactions of the Philological Society* 121(3). 546–567.

Wolk, Christoph, Bresnan, Joan, Rosenbach, Anette & Szmrecsanyi, Benedikt. 2013. Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. *Diachronica* 30(3). 382–419.

Zaenen, Annie & Jean Carlette & Gregory Garretson & Joan Bresnan & Andrew Koontz-Garboden & Tatiana Nikitina, Catherine O'Connor & Tom Wasow. 2004. *Animacy encoding in English: why and how*. Proceedings of the 2004 ACL workshop on discourse annotation. Barcelona. July 2004.

CONTACT

sali.tagliamonte@utoronto.ca

Linguistic Typology at the Crossroads 5-2 (2025): 188-223 ISSN 2785-0943

Language variety as a linguistic subsystem: typological implications

ALESSANDRO VIETTI[1], MASSIMO CERRUTI[2]
[1]FREE UNIVERSITY OF BOZEN-BOLZANO, [2]UNIVERSITY OF TURIN

Submitted: 14/11/2024 Revised version: 9/9/2025
Accepted: 10/9/2025 Published: 25/02/2026

Abstract

This paper emphasizes the importance of language variety in typological linguistics and its role in understanding inter- and intralingual variation. Adopting a corpus-based perspective, the paper explores the covariation structure that links linguistic phenomena and social information, revealing distinct linguistic subsystems within a language. Two case studies of sociolinguistic variation in Italian illustrate how language varieties emerging from the data can carry micro-typological meaning. The first study empirically tests the co-existence of 'neo-standard Italian' and the traditional literary standard in speakers' usage, while the second is a phonological analysis of the allophony of /r/ in Italian spoken by a bilingual community in South Tyrol (Italy). The analysis indicates that the integration of language varieties into typological research allows for a more complex representation of the hierarchical structure of linguistic phenomena.

Keywords: language variety; corpus-based typology; multivariate statistical analysis; Italian.

1. Introduction

Central to this paper is the notion of language variety, which is a key concept in variationist sociolinguistics (especially in European sociolinguistics; cf. Auer 1997; Guy & Hinskens 2016; Berruto 2004, 2019). A language variety is characterized by a set of co-occurring linguistic features, which are associated with language-external

factors, such as speakers' geographical provenance and social identity, or situational characteristics (cf. Milroy & Milroy 1997; Berruto 2009, 2012; Guy 2013).[1] From this perspective, therefore, a language is conceived as a diasystem (in the sense of Weinreich 1954); that is, a linguistic system coherently organized in regional, social and situational varieties. Such "heterogeneous system is then viewed as a set of subsystems which alternate according to one set of co-occurring rules" (Weinreich et al. 1968: 108).

Each language variety is indeed characterized by its own set of grammar rules. Every linguistic feature of a given variety is a "variant of a linguistic variable" (in Labov's 1966b classic definition) and results from the application of a variable rule, which is determined by the interplay of language-external and language-internal factors (cf. Sankoff 2004). A set of variable rules (with associated probabilistic weights) makes a language variety a linguistic subsystem[2] — that is, a grammatical system operating within the overall grammar of a language. This concept aligns with the notions of *Varietätengrammatik*, or 'variety grammar' (Klein 2004) and 'linguistic coherence' (Hinskens & Guy 2016; Beaman & Guy 2022).

In what follows, we will discuss how sociolinguistic investigations based on the notion of language variety can complement the analysis of cross-linguistic variation in typological research. First (Section 2), we will briefly outline two overall approaches to linguistic typology, paying special attention to the one that relies on quantitative corpus-based analysis and takes into account language-internal variation. Attention will then turn to two quantitative studies on spoken Italian[3] (Sections 3 and 4), both revolving around the detection of co-occurrence relationships between linguistic features and social variables that can imply typological relevance. In this paper, we refer to the analysis of these two published studies to propose a new interpretation of the results that emphasizes their typological implications. Finally (Section 5), based on the results of these studies, we will argue that the analysis of language-internal variation in a typological perspective cannot overlook the actual partitioning of a linguistic system into language varieties.

[1] From a different angle, a language variety or a social dialect can only be identified on the basis of a speech community ("the fundamental concept for identifying varieties is the language community"; Guy et al. 2022: 53), as defined by Labov (1966a; but see also Hymes 1967; Gumperz 1968).

[2] "Most of contemporary linguistics conceptualizes language varieties in terms of underlying linguistic systems – a grammar that is shared by the speakers of the variety" (Guy et al. 2022: 54).

[3] ITA; Indo-European, Romance.

2. Linguistic typology and variationist sociolinguistics

As far as typological research is concerned, the traditional approach can be contrasted with a more recent one. The former can be referred to as "grammar-based" (cf. Schnell & Schiborr 2022), or even "type-based" (cf. Levshina 2019). This line of research mainly draws on data coming from reference grammars and other descriptive accounts, such as linguistic features collated in large typological databases (e.g., WALS, Dryer & Haspelmath 2013). Accordingly, studies in this perspective generally rely on qualitative methods and are mostly based on linguistic features with allegedly categorical presence (even though typological databases show dominant types rather than categorical features; see, e.g., Dryer & Haspelmath 2013). Language-internal variation is, therefore, generally disregarded (Wälchli 2009).

Conversely, a more recent approach to linguistic typology can be referred to as "corpus-based" (cf. Levshina 2022; Schnell & Schiborr 2022), or even "token-based" (cf. Levshina 2019; see also Haspelmath 2018). This strand of research mainly draws on data coming from parallel and comparable linguistic corpora, or better yet from "language use, as approximated by corpora" (Levshina 2019: 534). Studies in this perspective are conducted with quantitative methods and employ continuous measures, such as the frequency or probability of occurrence (see, e.g., Gerdes et al. 2021) and probability-derived measures (e.g. Shannon entropy; cf. Levshina 2019), to investigate typologically relevant gradient properties of languages (e.g. word order variability). Linguistic features with variable presence clearly reflect language-internal variation; it is no coincidence that an increasing number of corpus-based typological studies give an account of both cross-linguistic and intra-linguistic variability (cf. Schnell & Schiborr 2022).

In sociolinguistic research, the concept of language variety has always been rather controversial. From the very beginning, in fact, it has been considered by some as a mere theoretical construct, unable to represent the actual behavior of speakers[4] (see, e.g., Hudson 1980). However, recent research has convincingly shown that the notion of language variety is firmly rooted in empirical evidence. Quantitative corpus-based studies conducted on actual speech data have indeed allowed for the detection of statistically robust sets of co-occurring features, each associated with socially-defined

[4] This issue has then come to the forefront with the advent of 'superdiverse' and globalized societies, which has challenged the explanatory power of most of the traditional notions in sociolinguistics (see, e.g., Blommaert 2010).

groups or communicative situations (see, e.g., Ghyselen & De Vogelaer 2018; Vietti 2019; Villena-Ponsoda & Vida Castro 2020; Beaman & Sering 2022).

Studies like these have affinities with corpus-based typological studies, since they both draw on data coming from corpora[5] and rely on quantitative methods. Nevertheless, quantitative measures employed by corpus-based typology tend to reduce language-internal variation to the presence of linguistic variants in different proportions. These measures can determine how much variation is found in a given dataset but disregard the arrangement of variants in a diasystem. Significantly, in typological studies that use corpus-based indices, each language (see, e.g., Levshina 2019, 2022; Schnell & Schiborr 2022: 176–177) or language variety (cf. Szmrecsanyi 2009) is assigned the value of an index of variability and regarded as a single data point. Languages or language varieties are then compared in terms of their respective index values.

From a sociolinguistic perspective, however, it is worth considering that every linguistic variant that comes into play in calculating the aforementioned indices is actually one of the co-occurring features that constitute a language variety; in other words, it is one of the structural units that form a grammatical subsystem (cf. Section 1). The simple use of indices of variability bears the risk of overlooking such an inner structure. A sociolinguistic approach can, therefore, complement corpus-based typology by providing interpretative tools for the analysis of the structure of language-internal variation. The following case studies provide us with an opportunity to elaborate on this issue.

3. Standard varieties in speakers' usage

The first study at hand (Cerruti & Vietti 2022) falls within the line of research dealing with the emergence of a new standard variety of Italian, the so-called 'neo-standard Italian'[6] (cf. Berruto 1987, 2012; Cerruti et al. 2017; Ballarè 2020), and aims to

[5] However, "it must be remarked that the use of corpora for typology is practically and methodologically more challenging than for sociolinguistics. The main issue is that cross-linguistic corpus-based studies need a reasonable number of corpora of different languages, which additionally must be sufficiently similar to allow a meaningful comparison" (Ballarè & Inglese 2023: 10; see also Levshina 2022).

[6] In fact, the convergence of the standard usage towards sub-standard varieties has led to the emergence of a new standard. It consists of (i) a nationwide shared core of originally sub-standard features that have come to be used and accepted even in formal and educated speech, as well as in writing, and (ii)

empirically test the co-existence of this newly emerged standard and the traditional, literary standard variety of Italian in speakers' usage.

3.1. Communicative situations and communicative purposes

This study, which considers the speech of university students as a case in point, is based on a subset of data coming from KIP (Mauri et al. 2019); i.e., a speech corpus consisting of different types of communicative interactions recorded at the universities of Turin and Bologna (two major cities in Northern Italy). In particular, a comparison is made between three types of interactions collected in Bologna: (a) spontaneous conversations among students recorded by in-group members; (b) semi-structured interviews collected by students within their peer-group; and (c) student-professor interactions during oral examinations.

This investigation takes into account 36 morphological, syntactical and lexical features; 23 of which are associated in the research literature with either neo-standard Italian or literary standard Italian, while the other 13 are related to the situational characteristics and/or the social attributes of speakers. As for the whole set of investigated features, the reader is referred to Cerruti & Vietti (2022: 265–267). Here we will focus on the 23 features associated with either neo-standard Italian or literary standard Italian (see Table 1).

	Label	**Feature**	**Standardness**
1.	Cleft	cleft sentence	neo-standard
2.	R_disloc	right dislocation	neo-standard
3.	L_disloc	left dislocation	neo-standard
4.	RET/GAP	pronoun retention/gap strategy (IO, OBL, GEN)	neo-standard
5.	MFNC.*che*	multifunctional *che*	neo-standard
6.	Dem Loc	demonstrative + locative adverb	neo-standard
7.	Dem *st**	aphaeretic forms of proximal demonstratives	neo-standard
8.	n_Dem *qu**	distance-neutral demonstrative *quello*(*/questo*)	neo-standard
9.	Cl *gli*	third-person singular indirect object clitic *gli* used for female referents	neo-standard

and a number of regionally marked features that have become part of the standard usage in distinct geographic areas. This is actually part of a more general process, which in recent years has led to the emergence of new standard varieties in many European languages (see, e.g., Kristiansen & Coupland 2011).

	Label	Feature	Standardness
10.	Refl 3	third-person singular and third person plural pronouns (*lui/lei* and *loro*) used as reflexive pronouns	neo-standard
11.	Loc *ci*	existential construction with *ci* as pre-copular proform	neo-standard
12.	*quell* che*	routinized formula *quello/i/a/e che*	neo-standard
13.	*tipo*	non-nominal uses of *tipo*	neo-standard
14.	Phrasal.V	phrasal verbs	neo-standard
15.	V*ci*	idiomatic verb-*ci* constructions	neo-standard
16.	V*ne*	idiomatic verb-*ne* constructions	neo-standard
17.	Pass	passive construction	literary standard
18.	REL	relative pronoun (IO, OBL, GEN)	literary standard
19.	Dem *qu**	*questo*-type proximal demonstratives, *quello*-type distal demonstratives	literary standard
20.	n_Dem *ciò*	distance-neutral demonstrative *ciò*	literary standard
21.	Cl *le*	third-person singular indirect object clitic *le* used for female referents	literary standard
22.	Refl *sé*	third person singular and third person plural reflexive pronoun *sé*	literary standard
23.	Loc *vi*	existential construction with *vi* as pre-copular proform	literary standard

Table 1: Neo-standard and literary standard features.

The main research question that the study seeks to answer is whether neo-standard Italian and literary standard Italian actually consist of two distinct sets of co-occurring linguistic features in the usage of speakers. Groups of features are detected by means of a Principal Component Analysis (PCA).[7] PCA results in a geometric representation of elements on a plane. In Figure 1, each point on the plane stands for a linguistic feature under scrutiny (features are labelled as in Table 1). The distance between the points represents the degree of correlation; that is, the nearer two points are on the plane, the stronger the co-occurrence between the two corresponding features in speakers' usage. Two bundles of features can thus be identified: a particularly cohesive group of features on the left, and a set of more broadly distributed features on the right.

[7] The use of PCA in the study of language varieties is well grounded in sociolinguistic research. The seminal application of this method can be found in Horvath & Sankoff (1987). Other highly significant works are cited in Walker, Hoffman & Meyerhoff (2022: 71–72) and Beaman & Sering (2022: 88–89).

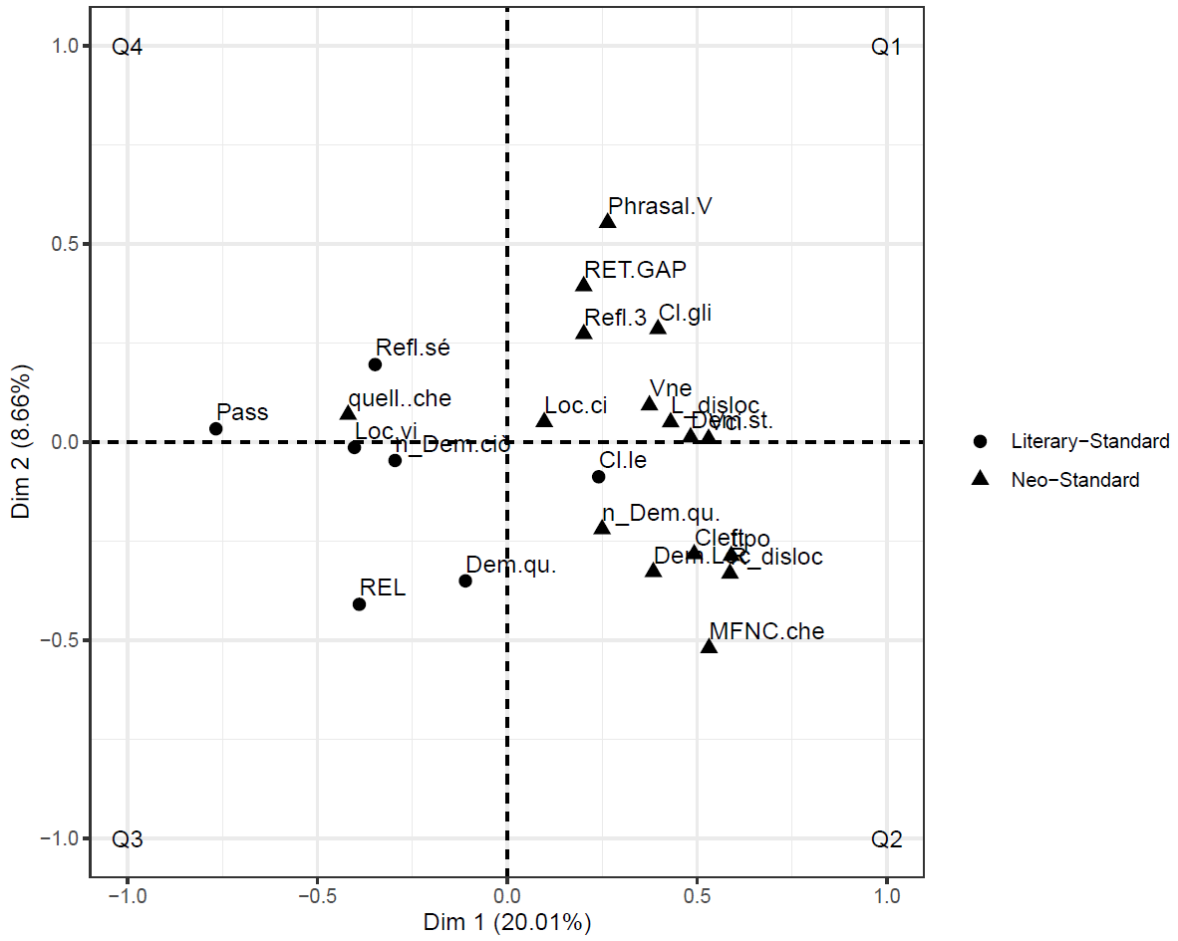


Figure 1: Plot of neo-standard and literary standard features in students' speech (Cerruti & Vietti 2022).

In Figure 2, each point represents a specific communicative interaction; the nearer two points are on the plane, the more features the corresponding communicative interactions have in common. In this respect, oral examinations stand apart from communications between peers (be they conversations or interviews). Figure 1 and Figure 2 can overlap. In fact, the linguistic features on the left half of Figure 1 tend to co-occur during oral examinations (left half of Figure 2), while the features on the right half of Figure 1 tend to co-occur during peer-to-peer communications (right half of Figure 2).

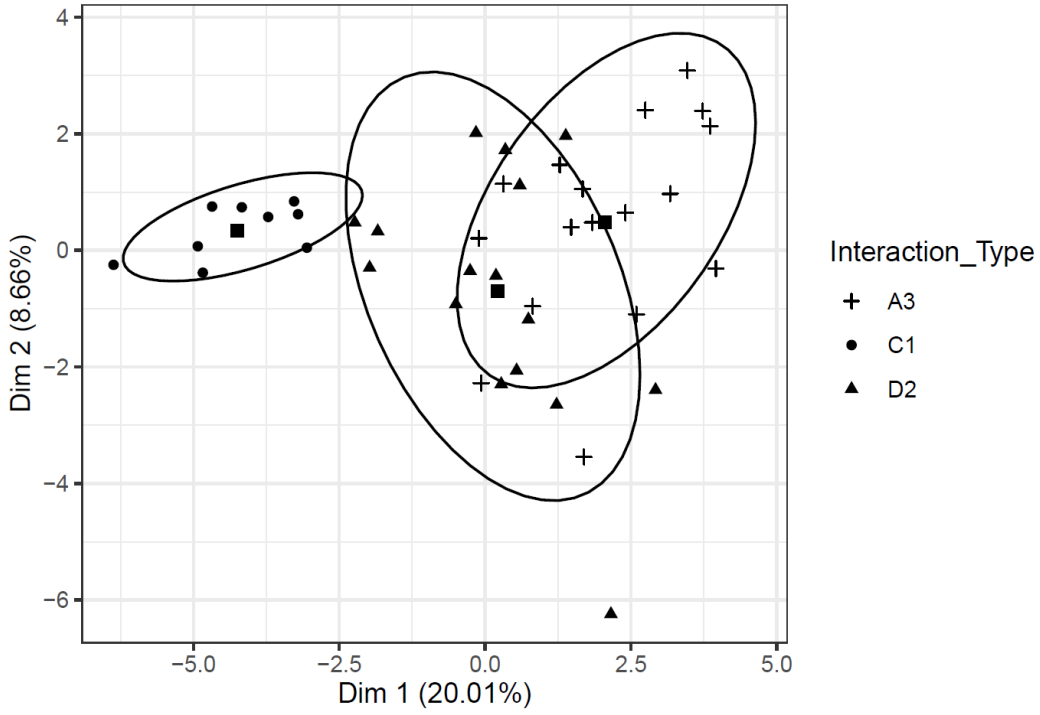


Figure 2: Plot of communicative interactions: spontaneous conversations (A3), semi-structured interviews (D2), oral examinations (C1) (Cerruti & Vietti 2022).

Moreover, in both Figure 1 and Figure 2 correlation patterns reflect the interplay of two underlying factors, represented by Dimension 1 (the horizontal axis) and Dimension 2 (the vertical axis). Dimension 1 can be associated with the opposition between formal and informal situations: the main difference between oral examinations (left half of Figure 2) and peer-to-peer communicative interactions (right half of Figure 2) relate to the asymmetrical vs. symmetrical relationship between the speakers; Dimension 2 can be associated with the opposition between interactional and informational purposes; the main differences between spontaneous conversations (upper end of Figure 2) and semi-structured interviews (lower end of Figure 2) indeed relate to the speakers' communicative intentions: participants in spontaneous conversations are engaged in expressing feelings, opinions, and attitudes, as well as indexing social identities, while participants in interviews are more focused on conveying information (cf. Biber & Conrad 2009).

In conclusion, PCA allows for the detection of two distinct clusters of co-occurring features (see Figure 1): one consisting of neo-standard features (with the exception of feature 21, Table 1) and the other consisting of literary standard features (with the exception of feature 12, Table 1). Neo-standard features tend to cluster during peer-to-peer communicative interactions, while literary standard features do so during oral

examinations. Two distinct language varieties can thus be identified in students' speech according to the symmetrical vs. asymmetrical relationship between the participants and, therefore, in accordance with the opposition between informality and formality.[8] At the same time, the co-occurrence of neo-standard features both in spontaneous conversations and in semi-structured interviews, namely in settings characterized by different communicative purposes, can reflect the multifunctional character of the newly emerged standard. Literary standard features, instead, seem to be more closely aligned with informational purposes.

3.2. Typological considerations

We now look at some new insights with these results. First of all, it is worth noting that most differences in grammar between literary standard Italian and neo-standard Italian are typologically relevant. In most cases, in fact, a given literary standard variant and its neo-standard counterpart represent two distinct values of "a structural property of language that describes one aspect of cross-linguistic diversity" (according to the definition of *feature* given in WALS, cf. Dryer & Haspelmath 2013); namely, two different values of a structural property of language that contribute in distinguishing between languages of different types (see also Comrie et al. 2013).

As regards, for example, relativization strategies, and in particular relativization on obliques, the literary standard use of a pronominal element signaling the syntactic role of the relativized item, as in utterance (1), can be contrasted with the neo-standard use of a clause-initial invariable complementizer, as in (2) (see also Cerruti 2017). This distinction clearly mirrors the opposition between two relativization strategies that can be found cross-linguistically (cf. Comrie & Kuteva 2013); that is, relative pronoun strategy and so-called 'gap' strategy (the latter involving cases in which the relative construction does not express the syntactic role of the relativized item).[9]

[8] The co-occurrence of literary standard features during oral examinations appears to reflect the role that the old standard continues to play in school education (cf. Berruto 2017). The presence of neo-standard features in other formal contexts, such as spoken and, to a certain extent, written media, is, however, beyond dispute, as is widely documented by an extensive body of research (see Ballarè 2020 for an up-to-date review).

[9] Another relativization strategy identified by Comrie & Kuteva (2013) is found to occur in our data, namely, pronoun retention; e.g., *il ragazzo che gli ho scritto una lettera* 'the boy to whom I wrote a letter', lit. 'the boy that to him I wrote a letter'. Positions lower than direct object on Keenan & Comrie's

(1) literary standard
la ragazza *con* *cui* *lavoravo* *di solito*
the girl with REL work.PST.IPFV.1SG usually
'the girl I used to work with'
(KIP Corpus, BOD1011)

(2) neo-standard
gente *che* *non* *faccio* *conoscenza*
people COMP NEG make.PRS.1SG acquaintance
'people I'm not acquainted with'
(KIP Corpus, BOA3004)

Further examples can be found in the domain of adnominal demonstratives. As for the traditional, literary standard, a norm 'without adjectives' has been replacing a Tuscan-based norm (cf. Da Milano 2015). The former reflects a distance-oriented system expressing a two-way contrast, in which a proximal demonstrative (*questo*) and a distal one (*quello*) indicate the relative distance between the referent and the speaker; see, e.g., excerpt (3). The Tuscan-based norm reflects a person-oriented system expressing a three-way contrast, in which a third demonstrative (*codesto*) denotes a referent near the hearer; see, e.g., example (4).[10] The binary opposition between a proximal demonstrative (*questo*, sometimes in the apheretic form *sto*, as in excerpt (5)) and a distal one (*quello*) is also drawn on in neo-standard Italian. However, neo-standard is distinguished by the optional presence of a demonstrative preceding the noun and a locative adverb (*qui/qua*, *lì*, *là*) following the noun, as in example (6) (cf. Berruto 2012: 87–88). This is a distance-oriented system that can express a three-way contrast, in which *lì* indicates a referent nearer than *là* (cf. Benedetti & Ricca 2002). The order of noun and demonstrative and the distance contrasts of demonstratives are known to be sensitive to cross-linguistic diversity (cf. Dryer 2013a; Diessel 2013).

(1977) Accessibility Hierarchy are relativized by pronoun retention in 7.08% of cases (16/226), by gap in 9.73% (22/226) and by relative pronouns in 83.19% (188/226). Subject and object relativization exclusively employ the gap strategy. The indirect object is relativized by relative pronouns in 92% of cases (23/25) and by pronoun retention in 8% (2/25).

[10] In our data, *codesto* is only found among speakers coming from Tuscany, as is the case with (4).

(3) literary standard
questo lavoro [...] *quel testo*
this work that text
'this work, that text'
(KIP Corpus, BOC1002, BOC1004)

(4) literary standard
codesto piatto
that dish
'that dish (near the hearer)'
(KIP Corpus, BOA3003)

(5) neo-standard
'sto rumore
this noise
'this noise'
(KIP Corpus, BOA3021)

(6) neo-standard
questo ragazzo qui […] quel libro lì […]
this boy here that book there
quelle cose là
those things there
'this boy, that book, those things (over there)'
(KIP Corpus, BOA3013, BOA3011, BOA3012)

Another source of diversity in the world's languages is the grammatical expression of co-reference between the subject and a non-subject argument within the same clause (cf. Haspelmath 2023). In the third person, literary standard Italian makes a distinction between a series of independent non-reflexive personal pronouns (see, e.g., *lei* in utterance (7)) and the independent reflexive pronoun *sé* (see, e.g., (8)), the latter being used as both a singular and plural form that signals co-reference with the subject of the clause. Conversely, neo-standard Italian can also use the series of non-reflexive third-person pronouns to express co-reference with the subject (cf. Cordin 2001: 610–611), as in utterance (9).

(7) literary standard, neo-standard
lui *tirava* *la* *palla* *a* *lei*
he throw.PST.IPFV.3SG the ball to her
'he threw the ball to her'
(KIP Corpus, BOA3002)

(8) literary standard
i *beni* *che* *la* *sposa* *porta* *con* *sé*
the goods COMP the bride take.PRS.3SG with herself
'the goods that the bride brings with herself'
(KIP Corpus, BOD1011)

(9) neo-standard
non *lavorava* *per* *gli* *altri,*
NEG work.PST.IPFV.3SG for the others
lavorava *per* *lei*
work.PST.IPFV.3SG for her
'she didn't work for others, she worked for herself'
(KIP Corpus, BOA3013)

Moreover, literary standard Italian shows gender marking on independent non-reflexive third-person pronouns, in both singular and plural forms. As for the latter, *essi* is used for male referents and *esse* for female referents; see, e.g., excerpts (10) and (11). On the contrary, neo-standard Italian uses a single third-person plural form (*loro*), as in (12), and, therefore, has no overt gender distinction in independent third-person plural pronouns (cf. Berruto 2012: 83–84).[11] On a cross-linguistic level, different values of gender contrasts in independent personal pronouns contribute in distinguishing between languages of different types (cf. Siewierska 2013).[12]

[11] Gender marking on third-person plural pronouns, or lack thereof, is not among the linguistic features considered in Cerruti & Vietti (2022). However, in the subset of KIP data collected in Bologna, the use of *essi/esse* is found to occur only during oral examinations (cf. *www.kiparla.it*).

[12] In the world-wide context, gender contrast on third-person pronouns, in both singular and plural forms, is rarer than only on third-person plural pronouns (cf. Siewierska 2013). Similarly, relative pronoun strategy is rarer than gap strategy in both oblique and subject relativization (cf. Comrie & Kuteva 2013). Cases like these lead Grandi (2019) to hypothesize that literary standard Italian is more 'exotic' (in the sense of Dahl 1990) than neo-standard Italian.

(10) literary standard

la	*casa*	*che*	*essi*	*fecero*	*per*	*noi*
the	house	COMP	they.M	make.PST.PFV.3PL	for	us

'the house they built for us'
(KIP Corpus, BOD1006)

(11) literary standard

qualsiasi	*esse*	*siano*
whatever	they.F	be.PRS.SBJV.3PL

'whatever they may be'
(KIP Corpus, BOD1009)

(12) neo-standard

loro	*non*	*capirebbero*	*una*	*parola*
they	NEG	understand.PRS.COND.3PL	a	word

'they wouldn't understand a word'
(KIP Corpus, BOC1004)

Differences in grammar between literary standard Italian and neo-standard Italian can also be found in the coding of spatial relations in motion events. In fact, literary standard tends to express multiple semantic components of a motion event on a single verb. For example, the verb can encode (a) motion and path (see, e.g., *scendo* in (13)), with the possibility for manner to be expressed by phrases or clauses appearing in an adverbial function (e.g., the prepositional phrase *di corsa* 'in a rush', cf. Bernini 2010: 30–31, 33–35), as in verb-framed languages (cf., *inter alia*, Talmy 1985, 2000), or even (b) motion, path and manner (see, e.g., *ho infilato* in (14)), with the contribution of prefixes coding path (cf. Iacobini & Vergaro 2012), as in satellite-framed languages (cf. Beavers et al. 2010). On the contrary, neo-standard Italian is more prone to the use of phrasal verb constructions (cf. Berruto 2017: 42), in which some semantic components are coded on the verb and others on an accompanying particle, as in satellite-framed languages. This is, for instance, the case with phrasal verbs coding motion on the root and path on an adverb (see, e.g., *vado giù* in utterance (15)), or both motion and manner on the main verb and path on an adverb (see, e.g., *si lancia giù* in excerpt (16)).

(13) literary standard
scendo *a* *prendere* *le* *brioche*
descend.PRS.1SG to take the brioches
'I'll go down and buy the brioches'
(KIP Corpus, BOD2004)

(14) literary standard
ho *infilato* *il* *telefono* *in* *tasca*
have.PRS.1SG slip.PTCP the phone in pocket
'I slipped the phone in my pocket'
(KIP Corpus, BOA3021)

(15) neo-standard
ho *detto* *vado* *giù*
have.PRS.1SG say.PTCP go.PRS.1SG down
'I said I'll go down'
(KIP Corpus, BOA3003)

(16) neo-standard
così *il* *gatto* *non* *si* *lancia* *giù*
so the cat NEG self throw.PRS.1SG down
'so the cat can't jump down'
(KIP Corpus, BOA3013)

3.3 Language types and language varieties

The handful of structural properties exemplified above suffices to show that each set of standard features, i.e. that of literary standard Italian and that of neo-standard Italian, is related not only to extra-linguistic motivations, such as the degree of formality of the situation or communicative purposes (see Section 3.1), but also to linguistic factors. Two of these are the overt coding of grammatical oppositions and the analytic vs. synthetic marking of meaning. As for the former, literary standard appears to express more contrasting grammatical meanings than neo-standard. It is indicative that literary standard uses relative pronouns to provide overt indication about the syntactic role of a relativized item (feature 18, Table 1; cf. utterance (1)),

draws on a reflexive pronoun to signal co-reference with the subject of the clause (feature 22, Table 1; cf. (7), (8)), and shows gender distinction in independent third-person plural pronouns (cf. (10), (11)). On the contrary, neo-standard can encode all syntactic roles of a relativized item by an invariable complementizer (feature 4, Table 1; cf. (2)), resorts to non-reflexive personal pronouns also to convey a reflexive meaning (feature 10, Table 1; cf. (9)), and has no gender contrast in third-person plural pronouns (cf. (12)).

Moreover, literary standard appears to be more inclined to synthetic marking of meanings than neo-standard. For example, literary standard uses a single element, i.e. a relative pronoun, to mark subordination and overtly signal the syntactic role of a relativized item (feature 18, Table 1; cf. utterance (1))[13] and tends to encode multiple semantic components of a motion event on a single verb (cf. (13), (14)). Analytic marking is more common in neo-standard. This is, for instance, the case with phrasal verb constructions, which encode some semantic components of a motion event on the verb and others on an accompanying particle (feature 14, Table 1; cf. (15), (16)). Similar considerations can even apply to the presence of a demonstrative simultaneously before and after a noun (feature 6, Table 1; cf. (6)), at least if we assume that the demonstrative preceding the noun is (or tends to be) used as a marker of definiteness or specificity (see, e.g., Parenti 2001; cf. Diessel 2013) and the locative adverb following the noun indicates a distance contrast; see, e.g., *'sta stampante lì ha sempre dato dei problemi* ('that printer has always malfunctioned', lit. 'this printer there has always malfunctioned'; Cerruti 2009: 90), in which the prenominal demonstrative appears to be distance-neutral.

It can, therefore, be argued that literary standard features tend to cluster (see Figure 1, Section 3.1) not only as a result of situational formality and informational purposes, but also because of common grammatical properties, such as the preference for overt coding of oppositions and synthetic marking of meanings. Overt coding and synthetic marking indeed reflect the characteristics of explicitness and 'integration' (in the sense of Chafe 1982) which are cross-linguistically associated with written, formal, and educated varieties (see, e.g., Biber & Conrad 2009). Similarly, neo-standard features can be argued to cluster (cf. Figure 1) not only under the impulse of informal and multifunctional communication, but also because of shared

[13] Different is the case of pronoun retention strategy (e.g., *il ragazzo che gli ho scritto una lettera* 'the boy to whom I wrote a letter', lit. 'the boy that to him I wrote a letter'; footnote 9), which combines a subordinating conjunction with a case-marked resumptive element.

grammatical properties, such as the tendency to leave inferable distinctions uncoded and convey meanings analytically. In fact, these properties are typical of vernaculars and, more generally, spoken varieties (as well as contact varieties; cf. Szmrecsanyi 2009; Haspelmath & Michaelis 2017) and consistent with the nature of neo-standard features, which originate mainly from spoken, informal, and uneducated varieties (see, e.g., Berruto 2012: 73–75).[14]

Leaving aside generalizations about intra-linguistic variation (cf. Kortmann 2004), it is worth noting that different typological profiles can coexist in the same diasystem and, most importantly, can take the form of different language varieties.

As shown in Grandi (2019, 2022; see also Ballarè & Inglese 2023), there are, in fact, significant similarities between the notions of language type and language variety. First and foremost, both notions are defined in terms of a correlation between linguistic features, steered by one or more factors. A language type can indeed be understood as a correlation between 'values of structural properties' (to use WALS terminology, cf. Section 3.2), which is determined by language-internal factors generally reflecting basic underlying parameters of linguistic structuring (e.g. locus of marking, alignment, head-directionality, etc.). For example, the head-directionality parameter appears to determine some recurrent correlations across languages, such as that between OV (Object-Verb), RelN (Relative-Noun) and NAdp (Noun-Adposition), on the one hand, and that between VO, NRel and AdpN, on the other hand (cf. Dryer 2013b, 2013c).

Similarly, a language variety is defined as a correlation between variants of linguistic variables (cf. Section 1) that is associated with language-external factors, such as the geographical origin and the social identity of the speakers, or situational characteristics. For example (with special reference to our case study), the opposition between formal and informal situations can be seen as underpinning some sub-systemic correlations, such as that between relative pronoun strategy, gender distinction in third-person plural pronouns and single verb constructions (as found in literary standard Italian), on the one hand, and that between gap strategy, the lack of gender marking in third-person plural pronouns and phrasal verb constructions (as found in neo-standard Italian), on the other hand.

[14] Needless to say, this is not at odds with the status of neo-standard features. They came to be used and accepted even in formal and educated speech, and to a certain extent in formal and educated writing (cf. footnote 6), as a result of a bottom-up process of (re)standardization (cf. Berruto 1987, 2012; Cerruti et al. 2017).

At the same time, the correlation between features that defines a language variety can be fostered not only by language-external factors but also by language-internal motivations. As argued above, for example, relative pronoun strategy tends to cluster with single verb constructions also due to the preference (that formal varieties show) for the synthetic marking of meanings, while gap strategy tends to cluster with the lack of gender contrast in third-person plural pronouns also because of the tendency (of informal varieties) to leave inferable distinctions uncoded. Moreover, as shown in Section 3.2, some of the co-occurring features that define a language variety represent a bundle of values of typologically relevant properties of language. This further establishes the notion of language variety as a suitable unit of analysis for both variationist sociolinguistics and typological research.

Finally, even the correlation between features that defines a language type can possibly be associated with both language-external and language-internal factors. This is particularly tenable when referring to areal types (cf. Grandi 2022: 139–140); i.e., groups of languages (such as those of the *Balkan Sprachbund*) that share a common set of linguistic features as a result of geographical proximity and language contact, rather than genetic relationships. One cannot fail to notice, moreover, that the traditional approach to linguistic typology (cf. Section 2) tends to ascribe a standard language to the type representing the standard variety of that language; in fact, the so-called "grammar-based" (or "type-based") typology mainly draws on data coming from reference grammars and thus focuses on the set of linguistic features that is codified as the reference model for the standard usage of a language. From this perspective, therefore, even a language type can be argued to have its own sociolinguistic, language-external, *côté*.

4. Phonetic variation in a bilingual community

The second study this paper considers is a phonological analysis of the realization of /r/ in Italian spoken by a bilingual community in South Tyrol (Italy). In this area, Italian and German, more specifically the Tyrolean dialect,[15] have been in stable contact for almost a century, a process that has led to the development of a contact variety of Italian, strongly influenced in its phonology by German. Together with the contact variety, the linguistic repertoire includes a regional variety, largely similar to

[15] Tyrolean is a German dialect belonging to the Southern Bavarian group (Wiesinger 1990).

the surrounding ones (such as Trentino and Veneto Italian; Vietti 2017; Vietti & Mereu 2023), spoken mainly by monolingual or Italian-dominant bilingual speakers.

Figure 3: Bolzano-Bozen (South Tyrol, Italy) and the two neighboring Italian regions (Trentino and Veneto).

Without partitioning the data into distinct varieties, the high variability in the production of /r/, in this context of language contact, would result in a redundant and unlikely system of allophonic relations among 11 variants distinguished by two constriction locations and four manners of articulation (see Table 2). The phonetic variability of /r/ poses, on a general level, a problem of phonetic-phonological characterization of this class of sounds (Lindau 1985; Wiese 2011), but in this particular case, the extreme mutability of realization raises a specific issue of mapping between allophones and contexts. On the one hand, the high variability in the production of /r/ (especially if uvular) does not easily allow for distinguishing between allophony and coarticulation (Iskarous et al. 2012); on the other hand, the frequent shift and contact between languages in a bilingual setting increases the entropy of the relationship between allophones and phonetic contexts: high entropy means many allophones for one phonetic context vs. low entropy where ideally the relationship is one allophone for one context (cf. Vietti & Spreafico 2018).

An analysis of how allophones co-occur in patterns associated with the social attributes of the speakers makes it possible to observe the self-organization of rhotic sounds into coherent subsystems traceable to distinct varieties of Italian. Instead of

looking at the frequency of the feature alone, this case study demonstrates the need to consider the distribution of a grammatical feature within a speech community to reveal its structuring into subsystems (Scobbie 2006; Docherty & Foulkes 2014).

4.1 The structure of /r/ allophonic variation

The dataset used in this study is based on reading and spontaneous speech tasks obtained from a sample of 14 speakers who were born and lived in Bolzano (South Tyrol).[16] The speakers are aged 24 and 38 years, with 9 female participants and 5 males). Among these speakers, 4 are categorized as sequential bilinguals with Italian as L1 and 8 as sequential bilinguals with Tyrolean as L1, while the remaining two participants are simultaneous bilinguals[17] (Paradis 2007).

In this corpus, we identified a set of 9 distinct realizations[18] of rhotic sounds, plus r-vocalization and deletion (see Table 2, Vietti & Spreafico 2016) that total 2276 tokens.

	Alveolar	**Retroflex**	**Uvular**
Trill	r		ʀ
Tap/Flap	ɾ	ɽ	ʀ̆
Approximant	ɹ		ʁ̞
Fricative	ɹ̝		χ ʁ

Table 2: Variants of /r/ in Bolzano Italian.

The primary goal of this analysis is to determine whether all /r/ variants pattern together in a single system of allophonic relations, however redundant, or whether variants are distributed among groups of speakers yielding distinct subsystems.

To observe patterns of association between allophones of /r/, phonetic contexts, and speakers, we used Correspondence Analysis (CA) for bivariate analysis and Multiple Correspondence Analysis (MCA) for multivariate analysis, two statistical

[16] The analysis is essentially similar to Vietti & Spreafico (2016), except for an additional correspondence analysis (Variants of /r/ - First Language) and, above all, the typological interpretation of the results.

[17] The speakers were classified on the basis of a questionnaire in which they self-assessed their linguistic background (see Kaland et al. 2016, Kaland et al. 2019).

[18] The variants of /r/ were identified through qualitative analysis of the spectrographic representations of the segments in Galatà, Vietti & Spreafico (2016).

techniques that are functionally similar to PCA (cf. Section 1) but differ in their mathematical treatment of variance and apply specifically to categorical variables[19] (Levshina 2015; Vietti & Spreafico 2016). The associations among values of the variables are expressed as chi-square distances and are represented graphically as points on a plot, uncovering the underlying structure of the data. Although the method of calculation and the metric by which correlations are expressed are different, the interpretation is similar to a PCA plot: the closeness between points is an indication of the strength of the association between categories of a variable. Similarly, the order of dimensions in MCA corresponds to the capacity to describe the variability of the data (variance in PCA, inertia in MCA). In this sense the first dimension is therefore the most important one.

To better understand the distinctive effect of allophone distribution across speakers, we performed a two-step analysis. In a first step, using Correspondence Analysis we observe associative patterns only among two linguistic variables (*Variants of /r/* and *Phonetic contexts*), simulating the existence of a single rhotic system of allophonic relations for Italian in South Tyrol. Variants of /r/ are coded with a labeling scheme in which the place precedes the manner of articulation, such as U_Trill (uvular trill) or A_Fric (alveolar fricative). Retroflex, vocalized /r/ and deletions are labeled as Retro, Vow and Del. Phonetic contexts analyzed are: #RV (initial syllable and word position as in *rosso* 'red'), CRV (consonant cluster in syllable onset as in *fretta* 'hurry'), VR# (final syllable and word position as in *bar* 'bar, café'), VR.C (syllable coda followed by a consonant as in *forchetta* 'fork'), VRRV (geminate /r/ as in *carro* 'cart'), VRV (syllable onset in intervocalic position as in *cuore* 'heart').

In the second step, we introduce a non-linguistic predictor into our analysis, i.e. *First language* in the acquisition process. This variable divides the possible paths of bilingual acquisition into three categories: Italian (sequential bilinguals:[20] Italian L1

[19] PCA aims to find new uncorrelated axes (principal components) that capture the maximum variance in the data. It is based on Euclidean distance between data points: the core idea is that if two data points are far apart in terms of their Euclidean distance, they are considered very different. MCA analyzes the "profiles" of individuals or categories. A "profile" is the set of relative frequencies of an individual's responses across all categories, or a category's prevalence across all individuals. MCA aims to find dimensions that highlight the greatest differences in these profiles. To do this, MCA uses a chi-squared distance metric that is sensitive to the relative differences between categories, not their absolute counts.

[20] Sequential bilinguals in South Tyrol acquire their second language around age 4-5, often at a pre-school stage.

– Tyrolean L2), German (sequential bilinguals: Tyrolean German L1 – Italian L2), and Italian-German (simultaneous bilinguals: Italian L1 + Tyrolean L1) (Paradis 2007). The second analysis aims at testing the groupings of linguistic variables by type of bilingual speaker, thus bringing out distinct and mutually incompatible allophonic subsystems, one related to the German-Italian contact variety and one to the regional variety.

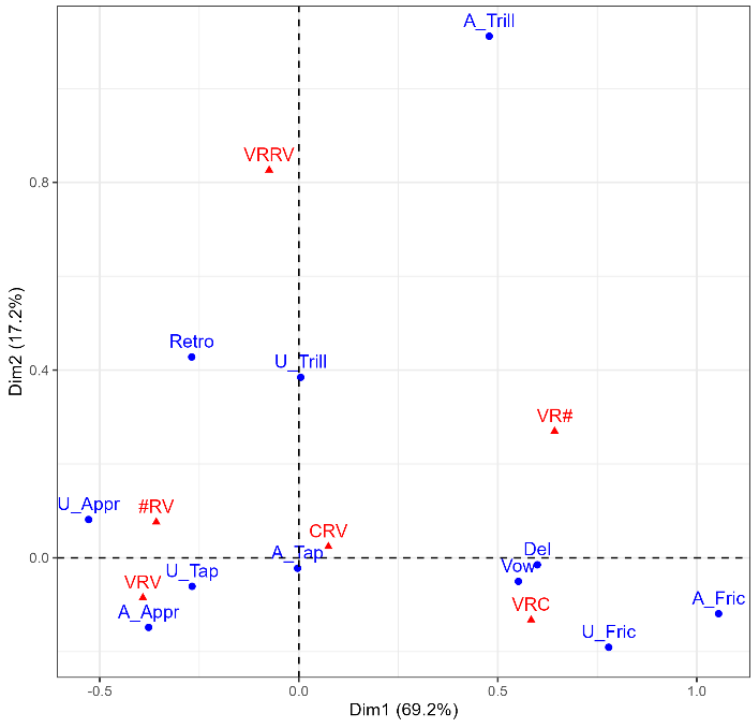


Figure 4: Plot of linguistic variants: /r/ variant (blue), Phonetic context (red).

The plot in Figure 4 reveals a general contextual distribution of allophones according to the manner of articulation (MoA). Dimension 1 distinguishes fricatives (on the right) from approximants (on the left) based on position within the syllable (coda vs. onset). Rhotic vocalization and deletions are found on the side of the syllabic coda contexts, while the taps (uvular and alveolar) are in a central position, closer to the origin of the plane, meaning that they are less context-dependent variants. Dimension 2 is mainly responsible for the distinction between the two trills and the retroflex and all other variants.[21] The two trills are weakly associated with VRRV, a phonetic

[21] The retroflex is a low-frequency phonetic realization in the corpus, the presence of which can be traced back to the influence of Veneto Italian on the Bolzano variety (Canepari 1986; Vietti 2017).

context in which the consonantal length is expressed as contrastive in Italian (Bertinetto & Loporcaro 2005; Krämer 2009). In the phonemic system of Italian spoken in Bolzano, the consonantal length distinction is not encoded simply as a gemination of the same segment (as, for example, for /l/ in *pala* [ˈpaːla] 'shovel' vs. *palla* [ˈpalla] 'ball'), but through two distinct phonetic realizations: tap as the default variant and trill for word-internal geminate contexts.

In this first analysis, which simulates a single allophonic system, two variants similar in manner but distinct in place of articulation may occur in the same phonetic context. This results in a highly redundant system in which sounds are potentially contrastive because the two conditions of phonemicity seem to be met: both the distribution of allophones by manner is unpredictable, and the variants are sufficiently distinct from a perceptual point of view (Renwick & Ladd 2016; Kiparsky 2015). Although the phonological conditions for phonemicity are satisfied, the lack of lexical distinctiveness does not establish an opposition with place of articulation between MoA rhotic pairs. In addition, the hypothetical system that emerges from this distributional analysis would also be very unlikely from a typological point of view, as consonant inventories containing two rhotics rarely contrast by place of articulation[22] (Maddieson 1988: 84–88). Rather, it seems, given the pressure of language contact, that we are dealing with a diasystem in which two subsystems are overlapping in phonetic and distributional properties (Pulgram 1964).

In the second plot (Figure 5), the introduction of correlations between linguistic variables and speakers' properties disentangles the apparent superposition of the subsystems by separating them into two groups of rhotic sounds. When the predictor *First Language* is added to the Multiple Correspondence Analysis, dimension 1 divides variants of /r/ by place of articulation: the uvular rhotics are found on the right side of the plane along with Tyrolean L1 sequential bilinguals (Ger) and simultaneous bilinguals (Ita-Ger), while on the left side, the alveolar rhotics are associated with Italian L1 sequential bilinguals (Ita).

[22] In some southern Swedish dialects, for example, dorsal and coronal /r/ are present in the inventory, but they are in complementary distribution (Engstrand et al. 2007).

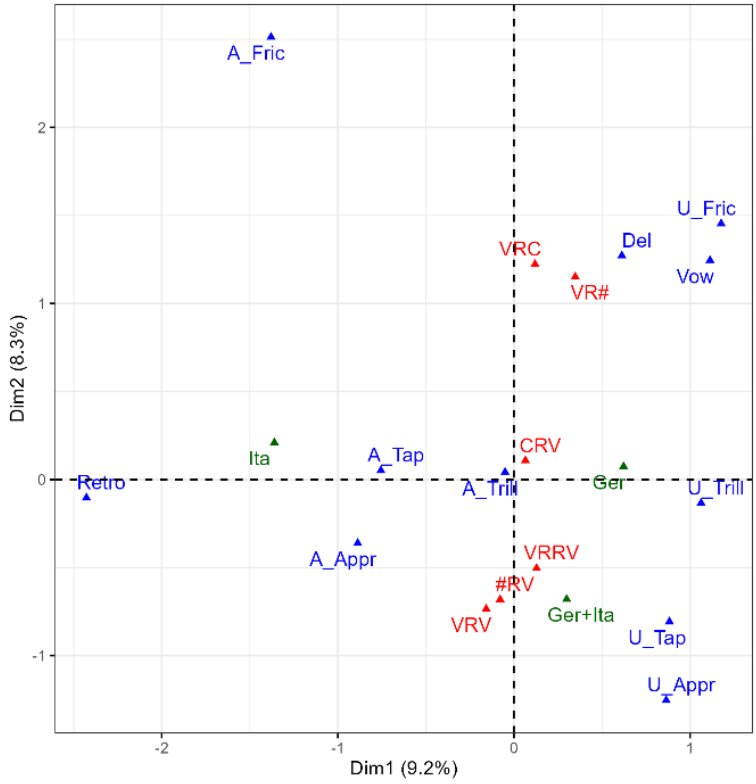


Figure 5: Plot of sociolinguistic variants: /r/ variant (blue), Phonetic context (red), First language of acquisition (green).

In this dimension, however, the two groups differ with respect to internal cohesion. The uvular rhotics, vocalizations and deletions are very close to each other and more clearly identify a German variety of Italian, due to the association with Tyrolean as L1 (Ger). In contrast, the alveolar rhotics show greater differences, e.g. retroflex takes on more extreme values as it correlates with one speaker's Veneto regional Italian,[23] while the alveolar trill has a central position with respect to the plot. Alveolar trill position is related to the fact that while L1 Italian speakers do not categorically use back variants of /r/, some simultaneous bilinguals (Ger+Ita) and Tyrolean L1 sequential bilinguals (Ger) may use front variants (e.g., the alveolar trill) when speaking Italian. The imperfect separation between variants by place of articulation and type of bilingual justifies the reduced ability of MCA to explain the variance in the data (9.2% and 8.3% for the first two dimensions). However, since no speaker

[23] As already noted in Vietti (2017), the Italian spoken in Bolzano is the result of a process of koineization between geographical varieties spoken in Italian regions bordering South Tyrol, such as Veneto (see Fig. 3). The Veneto component forms the basis of the Bolzano koine and today has taken on the social meaning of a broad local accent.

uses both variants of PoA in Italian, this means the subsystems likely cannot coexist in a single speaker's production and therefore that the integration of social information in the MCA is crucial to bring out the internal structure of the data.

If we consider Dimension 2, the variants seem to follow a rather weak pattern of association between phonetic contexts and manner of articulation. Pairs of /r/-sounds similar in manner are loosely mapped to the same phonetic contexts according to a pattern that seems to favor many-to-many relationships rather than a "one allophone-one context" relationship (Hall 2013). For example, the two fricative /r/ sounds (A_Fric and U_Fric) are located in the same upper half-plane and are weakly mapped to postvocalic /r/ settings (VRC and VR#), but the same is true for /r/ vocalizations and deletions. The introduction of the variable *First language* brings some order to the allophonic variation of /r/, but does not yet lead to the clear identification of coherent allophonic subsystems. One possible explanation for the increased variability and ambiguity of associations may be that the introduction of social information, on the one hand, allows for the emergence of two potential varieties of Italian, distinct by place of articulation, but, on the other hand, covers up incompatible allophone systems. For example, the alveolar trill is used by both Italian-L1 sequential bilinguals and simultaneous bilinguals (and also some Tyrolean-L1 sequential bilinguals) but arguably with different distributional rules. Differences between systems of allophonic relations are thus masked in a single flattened representation.[24] However, even with these limitations, the analysis reveals constraints on the high variability of /r/ that allow us to understand more clearly the nature of allophonic variation and the structure of the data.

The order manifests itself on two levels: the first concerns the partition of variation into two distinct subsystems on the basis of the social distribution of variants; the second shows the probabilistic nature of the relationships between allophones and phonetic contexts (Hall 2013), which is certainly amplified in a situation of language contact.

The analysis of the covariation bonds between linguistic and social features allows us to understand the relationship between the levels in which sociolinguistic reality

[24] Correspondence analysis on two variables (Variants of /r/ and First language), without considering the association with phonetic contexts, indicates a high variance explained by the first dimension (83.8%). This means that the social predictor is able to clearly order allophones by language variety, but also that within each language variety, the association between /r/ allophones and phonetic contexts does not generate fully coherent subsystems.

is organized into grammatical subsystems or varieties, which then is nothing more than the recognition of the orderly heterogeneity of language variation (Weinreich et al. 1968; Beaman & Guy 2022). The organization of a language into coexisting subsystems may or may not be reflective of variation at the (micro)typological level, but failure to recognize this multilevel, multidimensional articulation can lead to misinterpretation of the typological nature of a language system, whether under a grammar-based or frequency-based approach.

5. Conclusions

The aim of the paper has been to show the relevance of the notion of language variety for research in typological linguistics and, more broadly, to emphasize the continuity between variation across languages and their internal structuring. We have defined language varieties as minimally distinct grammatical subsystems within a larger language system, in which the links binding linguistic features emerge through co-occurrence with social units, such as the communicative situation or the speaker. As we have shown, the features of a language variety can also exhibit linguistic coherence not only for local communicative purpose, but also at the level of typological pattern in the expression of grammatical categories.

Underlying our analysis is the assumption that different approaches to typological research involve distinct interpretations of what should be considered a unit of linguistic observation. In type- or grammar-based typology, the unit of analysis is a descriptive account of a grammatical system, for which the linguistic feature to be analyzed will be of a discrete type and will take on a single categorical value for that grammar. For example, in the case of the typology of object and verb order, although internal variability is recognized, a dominant type will represent an entire grammar.

In corpus-based typology, the unit of analysis is a corpus and thus the (value of a) linguistic feature is represented as its frequency distribution. The typological variation of a linguistic feature can be analyzed in two ways, marginal and conditional, that is, univariate or bi- and multivariate. In the example of verb and object order, this means observing (a) the marginal frequency of the possible orders or (b) the conditional frequency, i.e. searching for dependencies between the frequency of orders and other relevant phenomena in the linguistic structure such as e.g. case marking or agreement. This second possibility can also be investigated more

abstractly with entropy measures that continuously capture the internal variability of a distribution (Koplenig et al. 2017; Levshina 2019).

According to this line of research, the reliability of inference relies on the representativeness of a corpus and comparability between corpora. The former ensures the generalizability of the inference, while the latter prevents uncontrolled variables from exerting undesirable effects on the variation of the observed linguistic feature. The study of any linguistic phenomenon from a corpus-based typology perspective is conducted on homogeneous datasets, i.e. based on the same type of text, so as to control for possible sources of covariation.

If comparing data from different languages requires a necessary selection of homogeneous datasets, what if corpora containing balanced sociolinguistic variation could be compared? What steps should be taken to make the correct crosslinguistic inferences from the data? In this paper, we aimed to focus on the covariation structure that links linguistic phenomena and social information and how the analysis of this structure can help identify distinct grammatical subsystems within an overall linguistic system. In particular, the diasystem of a language can be usefully decomposed into language varieties through the association between bundles of linguistic features and meaningful social units, such as the communicative situation and the social characteristics of the speaker. For this purpose, we employed statistical methods such as PCA and MCA that allow us to observe the emergence of patterns of association between variables.

The analysis conducted on these two cases of sociolinguistic variation in Italian show how the notion of language variety can have relevant implications for research in corpus-based typology. First, a corpus-based approach allows for the empirical identification of language varieties as emergent structures of covariation between linguistic and social elements. The first case study, for example, illustrates how the two identified standard varieties of Italian can coexist in speakers' competence to express stylistic differences along a scale of greater or lesser formality. Second, the varieties can also be linguistically coherent, that is, show how the socio-communicative purpose of usage can lead to the harmonic selection of categories that linguistic features can take on. As a result, the structuring of a language into relatively stable varieties makes micro-typological variation possible, as is the case with the different grammatical encoding strategies of the two standard varieties of Italian (Section 3).

Finally, varieties may also be subsystems related to groups of speakers and not necessarily to contexts of language use, which may be non-uniformly distributed in the speech community. This implies the presence of varieties that do not co-exist in the speakers' competence and therefore are potentially more dissimilar to each other. In the case of the study of allophonic variation of /r/ in contact varieties of Italian, failure to divide the micro-system of /r/ into two sets of variants would lead to a grammatical misdescription. In other words, a marginal analysis of the distribution of variants not conditioned by social variables would result in a phonemic category with anomalous allophonic variability, characterized by high contextual entropy, creating extreme difficulty in matching allophones and phonetic contexts.

Our study is limited in several respects, mainly because the two case studies are very specific in the linguistic features investigated and are also related to only one language, Italian. In addition, analyses of this kind are severely constrained methodologically by the nature of the source data and the difficulty in scaling crosslinguistically due to the lack of comparable and representative sociolinguistic corpora. However, we have illustrated how the concept of 'language variety', as a grammatical subsystem that can be identified through its social distribution, can be a useful heuristic in the corpus-based typology research program. Language variety plays a pivotal role within the hierarchical data structure that links individual linguistic features at one end to the entire population of sentence tokens (i.e., the language corpus) at the other end.

Abbreviations

1 = 1st person
3 = 3rd person
COMP = complementizer
COND = conditional
F = feminine
IPFV = imperfective
M = masculine
NEG = negative
PFV = perfective
PL = plural
PRS = present
PST = past
PTCP = participle
REL = relative
SBJV = subjunctive
SG = singular

References

Auer, Peter. 1997. Co-Occurrence Restrictions between Linguistic Variables: A Case for Social Dialectology, Phonological Theory and Variation Studies. In Hinskens,

Frans L. & Van Hout, Roeland & Wetzels, W. Leo (eds.), *Current Issues in Linguistic Theory*, vol. 146, 69. Amsterdam: John Benjamins Publishing Company. (doi:10.1075/cilt.146.05aue)

Ballarè, Silvia. 2020. L'italiano neo-standard oggi: stato dell'arte. *Italiano LinguaDue* 12(2). 469–492. (doi:10.13130/2037-3597/15013)

Ballarè, Silvia & Inglese, Guglielmo. 2023. Analyzing language variation: Where sociolinguistics and linguistic typology meet. In Ballarè, Silvia & Inglese, Guglielmo (eds.), *Sociolinguistic and Typological Perspectives on Language Variation*, 1–27. Berlin, Boston: De Gruyter.

Beaman, Karen V. & Sering, Konstantin. 2022. Measuring change in lectal coherence across real- and apparent-time. In Beaman, Karen & Guy, Gregory R. (eds.), *The coherence of linguistic communities: orderly heterogeneity and social meaning*, 87–105. New York: Routledge.

Beaman, Karen V. & Guy, Gregory R. (eds.). 2022. *The coherence of linguistic communities: orderly heterogeneity and social meaning* (Routledge Studies in Sociolinguistics). New York, NY: Routledge.

Beavers, John & Levin, Beth & Tham, Shiao Wei. 2010. The typology of motion expression revisited. *Journal of Linguistics* 46(3). 331–377.

Benedetti, Marina & Ricca, Davide. 2002. The systems of deictic place adverbs in the Mediterranean: Some general remarks. *Mediterranean languages. Papers from the MEDTYP workshop*, 13–32. Bochum: Brockmeyer.

Bernini, Giuliano. 2010. Word classes and the coding of spatial relations in motion events: A contrastive typological approach. In Marotta, Giovanna & Lenci, Alessandro & Meini, Linda & Rovai, Francesco (eds.), *Space in Language. Proceedings of the Pisa International Conference*, 29–52. Pisa: ETS.

Berruto, Gaetano. 1987. *Sociolinguistica dell'italiano contemporaneo*. Roma: La Nuova Italia Scientifica.

Berruto, Gaetano. 2004. Sprachvarietät - Sprache (Gesamtsprache, historische Sprache). . In Ammon, Ulrich & Dittmar, Norbert & Mattheier, Klaus J. & Trudgill, Peter (eds.), *Sociolinguistics/Soziolinguistik. Ein internationales Handbuch zur Wissenschaft von Sprache und Gesellschaft*, vol. 1, 188–195. 2nd edn. Berlin, New York: De Gruyter.

Berruto, Gaetano. 2009. Identifing dimensions of linguistic variation in a language space. In Auer, Peter & Schmidt, Jürgen Erich (eds.), *Language and Space. An*

International Handbook of Linguistic Variation. Vol. 1. Theories and Methods, 226–241. Berlin, New York: Mouton de Gruyter. (doi:10.1515/9783110220278.226)

Berruto, Gaetano. 2012. *Sociolinguistica dell'italiano contemporaneo*. Roma: Carocci.

Berruto, Gaetano. 2017. What is changing in Italian today? Phenomena of restandardization in syntax and morphology: an overview. In Cerruti, Massimo & Crocco, Claudia & Marzo, Stefania (eds.), *Towards a New Standard*, 31–60. Berlin, New York: De Gruyter. (doi:10.1515/9781614518839-002)

Berruto, Gaetano. 2019. La nozione di 'varietà di lingua': una categoria obsoleta? In Bidese, Ermenegildo & Casalicchio, Jan & Moroni, Manuela Caterina (eds.), *La linguistica vista dalle Alpi: teoria, lessicografia e multilinguismo: studi in onore di Patrizia Cordin = Linguistic views from the Alps: language theory, lexicography and multilingualism: studies in honor of Patrizia Cordin* (Studia Romanica et Linguistica 57), 213–236. Berlin: Peter Lang.

Bertinetto, PierMarco & Loporcaro, Michele. 2005. The sound pattern of Standard Italian, as compared with the varieties spoken in Florence, Milan and Rome. *Journal of the Phonetic Association* 35(2). 131–151.

Biber, Douglas & Conrad, Susan. 2009. *Register, Genre, and Style*. Cambridge: Cambridge University Press. (doi:10.1017/CBO9780511814358)

Blommaert, Jan. 2010. *The Sociolinguistics of Globalization*. 1st edn. Cambridge: Cambridge University Press. (doi:10.1017/CBO9780511845307)

Canepari, Luciano. (1986). *Lingua italiana nel Veneto*. CLESP.

Cerruti, Massimo. 2009. *Strutture dell'italiano regionale: morfosintassi di una varietà diatopica in prospettiva sociolinguistica* (Spazi comunicativi 7). Frankfurt am Main: Peter Lang.

Cerruti, Massimo. 2017. Changes from below, changes from above: relative constructions in contemporary Italian. In Cerruti, Massimo & Crocco, Claudia & Marzo, Stefania (eds.), *Towards a New Standard*, 61–88. Berlin, New York: De Gruyter. (doi:10.1515/9781614518839-003)

Cerruti, Massimo & Crocco, Claudia & Marzo, Stefania (eds.). 2017. *Towards a New Standard: Theoretical and Empirical Studies on the Restandardization of Italian*. De Gruyter. (doi:10.1515/9781614518839)

Cerruti, Massimo & Vietti, Alessandro. 2022. Identifying language varieties: Coexisting standards in spoken Italian. In Beaman, Karen V. & Guy, Gregory R. (eds.), *The Coherence of Linguistic Communities. Orderly Heterogeneity and Social Meaning* (Routledge Studies in Sociolinguistics), 262–280. New York: Routledge.

Chafe, Wallace L. 1982. Integration and Involvement in speaking, writing, and oral literature. In Tannen, Deborah (ed.), *Spoken and written language: Exploring orality and literacy*, 35–53. Norwood: Ablex.

Comrie, Bernard & Dryer, Matthew S. & Gil, David & Haspelmath, Martin. 2013. Introduction (v2020.3). *The world atlas of language structures (online)*. Zenodo. (Ed. Dryer, Matthew S. & Haspelmath, Martin.) (doi:10.5281/zenodo.7385533)

Comrie, Bernard & Kuteva, Tania. 2013. Relativization on Obliques. *The World Atlas of Language Structures Online (v2020.3)*. Zenodo. (Ed. Dryer, Matthew S. & Haspelmath, Martin.) (doi:10.5281/zenodo.7385533)

Cordin, Patrizia. 2001. I pronomi riflessivi. In Renzi, Lorenzo & Salvi, Giampaolo & Cardinaletti, Anna (ed.), *Grande grammatica italiana di consultazione*, vol. I, 607–617. Bologna: Il Mulino.

Da Milano, Federica. 2015. Italian. In Jungbluth, Konstanze & Da Milano, Federica (eds.), *Manual of Deixis in Romance Languages*, 59–74. De Gruyter. (doi:10.1515/9783110317732-006)

Dahl, Östen. 1990. Standard Average European as an Exotic Language. In Bechert, Johannes & Bernini, Giuliano & Buridant, Claude (eds.), *Toward a Typology of European Languages* (Empirical Approaches to Language Typology), 3–8. Berlin: Mouton de Gruyter.

Diessel, Holger. 2013. Distance Contrasts in Demonstratives. *The World Atlas of Language Structures Online (v2020.3)*. Zenodo. (Ed. Dryer, Matthew S. & Haspelmath, Martin.) (doi:10.5281/zenodo.7385533)

Docherty, Gerard J. & Foulkes, Paul. 2014. An evaluation of usage-based approaches to the modelling of sociophonetic variability. *Lingua* 142. 42–56. (doi:10.1016/j.lingua.2013.01.011)

Dryer, Matthew S. 2013a. Order of Demonstrative and Noun. *The World Atlas of Language Structures Online (v2020.3)*. Zenodo. (Ed. Dryer, Matthew S. & Haspelmath, Martin.) (doi:10.5281/zenodo.7385533)

Dryer, Matthew S. 2013b. Relationship between the Order of Object and Verb and the Order of Relative Clause and Noun. *The World Atlas of Language Structures Online (v2020.3)*. Zenodo. (Ed. Dryer, Matthew S. & Haspelmath, Martin.) (doi:10.5281/zenodo.7385533)

Dryer, Matthew S. 2013c. Relationship between the Order of Object and Verb and the Order of Adposition and Noun Phrase. *The World Atlas of Language Structures Online*

(v2020.3). Zenodo. (Ed. Dryer, Matthew S. & Haspelmath, Martin.) (doi:10.5281/zenodo.7385533)

Dryer, Matthew S. & Haspelmath, Martin (eds.). 2013. WALS Online (v2020.3). Zenodo. (doi:10.5281/zenodo.7385533)

Engstrand, Olle & Frid, Johan & Lindblom, Björn. 2007. A Perceptual Bridge Between Coronal and Dorsal /r/. In Solé, Maria-Josep & Beddor, Patrice Speeter & Ohala, Manjari (eds.), *Experimental Approaches to Phonology*, 175–191. Oxford: Oxford University Press. (doi:10.1093/oso/9780199296675.003.0012)

Galatà, Vincenzo & Spreafico, Lorenzo & Vietti, Alessandro & Kaland, Constantijn. 2016. An acoustic analysis of /r/ in Tyrolean. *INTERSPEECH 2016*, 1002–1006. San Francisco, USA. (doi:10.21437/Interspeech.2016-418)

Gerdes, Kim & Kahane, Sylvain & Chen, Xinying. 2021. Typometrics: From Implicational to Quantitative Universals in Word Order Typology. *Glossa: a journal of general linguistics*. Open Library of Humanities 6(1). (doi:10.5334/gjgl.764)

Ghyselen, Anne-Sophie & De Vogelaer, Gunther. 2018. Seeking Systematicity in Variation: Theoretical and Methodological Considerations on the "Variety" Concept. *Frontiers in Psychology* 9. (doi:10.3389/fpsyg.2018.00385)

Grandi, Nicola. 2019. Che tipo, l'italiano neostandard! In Moretti, Bruno & Kunz, Aline & Natale, Silvia & Krakenberger, Etna (eds.), *Le tendenze dell'italiano contemporaneo rivisitate*, 59–74. Milano: Officinaventuno.

Grandi, Nicola. 2022. La variazione tra le lingue e nelle lingue. In Grandi, Nicola & Mauri, Caterina (eds.), *La tipologia linguistica. Unità e diversità nelle lingue del mondo*, 121–173. Roma: Carocci.

Gumperz, John. 1968. The Speech Community. In Sills, David L. (ed.), *International Encyclopedia of the Social Sciences*, 381–386. London: Macmillan.

Guy, Gregory R. 2013. The Cognitive Coherence of Sociolects: How Do Speakers Handle Multiple Sociolinguistic Variables? *Journal of Pragmatics* 52. 63–71.

Guy, Gregory R. & Hinskens, Frans. 2016. Linguistic coherence: Systems, repertoires and speech communities. *Lingua* 172–173. 1–9. (doi:10.1016/j.lingua.2016.01.001)

Guy, Gregory R. & Oushiro, Livia & Mendes, Ronald Beline. 2022. Indexicality and coherence. In Beaman, Karen & Guy, Gregory R. (eds.), *The coherence of linguistic communities: orderly heterogeneity and social meaning*, 53-68. New York: Routledge.

Hall, Kathleen Currie. 2013. A typology of intermediate phonological relationships. *The Linguistic Review*. De Gruyter Mouton 30(2). 215–275. (doi:10.1515/tlr-2013-0008)

Haspelmath, Martin. 2018. How comparative concepts and descriptive linguistic categories are different. In Van Olmen, Daniël & Mortelmans, Tanja & Brisard, Frank (eds.), *Aspects of Linguistic Variation* (Trend in Linguistics 324), 83–114. Berlin, New York: Mouton de Gruyter. (doi:10.1515/9783110607963-004) (Accessed May 29, 2023.)

Haspelmath, Martin. 2023. Comparing reflexive constructions in the world's languages. In Janic, Katarzyna & Puddu, Nicoletta & Haspelmath, Martin (eds.), *Reflexive constructions in the world's languages* (Research on Comparative Grammar 3), 19–62. Berlin: Language Science Press.

Haspelmath, Martin & Michaelis, Susanne. 2017. Analytic and synthetic: Typological change in varieties of European languages. In Buchstaller, Isabelle & Siebenhaar, Beat (eds.), *Language variation – European perspectives VI: Selected papers from the 8th International Conference on Language Variation in Europe (ICLaVE 8), Leipzig 2015*, 3–22. Amsterdam: Benjamins.

Hinskens, Frans & Guy, Gregory R. (eds.). 2016. *Coherence, covariation and bricolage. Various approaches to the systematicity of language variation*. Thematic issue of *Lingua* 172/173.

Horvath, Barbara & Sankoff, David. 1987. Delimiting the Sydney Speech Community. *Language in Society* 16. 179–204.

Hudson, Richard Anthony. 1980. *Sociolinguistics* (Cambridge Textbooks in Linguistics). Cambridge: Cambridge University Press.

Hymes, Dell. 1967. Models of the interaction of language and social setting. *Journal of Social Issues* 23(2). 8–28.

Iacobini, Claudio & Vergaro, Carla. 2012. Manner of motion verbs in Italian: semantic distinctions and interlingual comparisons. In Ferreri, Silvana (ed.), *Lessico e lessicologia*, 71–87. Roma: Bulzoni.

Iskarous, Khalil & McDonough, Joyce & Whalen, D. H. 2012. A gestural account of the velar fricative in Navajo. *Laboratory Phonology* 3(1). 195–210. (doi:10.1515/lp-2012-0011)

Kaland, Constantijn & Galatà, Vincenzo & Spreafico, Lorenzo & Vietti, Alessandro. 2016. /r/ as language marker in bilingual speech production and perception.

INTERSPEECH 2016, 515–519. San Francisco, USA. (doi:10.21437/Interspeech.2016-418)

Kaland, Constantijn & Galatà, Vincenzo & Spreafico, Lorenzo & Vietti, Alessandro. 2019. Which Language R You Speaking? /r/ as a Language Marker in Tyrolean and Italian Bilinguals. *Language and Speech* 62(1). 137–163. (doi:10.1177/0023830917746551)

Keenan, Edward L. & Comrie, Bernard. 1977. Noun phrase accessibility and Universal Grammar. *Linguistic Inquiry* 8. 63–99.

Kiparsky, Paul. 2015. New perspectives in historical linguistics. In Bowern, Claire & Evans, Bethwyn (eds.), *The Routledge Handbook of Historical Linguistics*, 64–102. London, New York: Routledge.

Klein, Wolfgang. 2004. Varietätengrammatik. In Ammon, Ulrich & Dittmar, Norbert & Mattheier, Klaus J. (eds.), *Soziolinguistik. Ein internationales Handbuch zur Wissenschaft von Sprache und Gesellschaft*, 1163–1172. Berlin, New York: Mouton de Gruyter.

Koplenig, Alexander & Meyer, Peter & Wolfer, Sascha & Müller-Spitzer, Carolin. 2017. The statistical trade-off between word order and word structure – Large-scale evidence for the principle of least effort. *PLOS ONE* 12(3). e0173614. (Ed. Smith, Kenny.) (doi:10.1371/journal.pone.0173614)

Kortmann, Bernd (ed.). 2004. *Dialectology meets typology: dialect grammar from a cross-linguistic perspective* (Trends in Linguistics Studies and Monographs 153). Berlin: Mouton de Gruyter.

Krämer, Martin. 2009. *The phonology of Italian* (Oxford Linguistics). Oxford, New York: Oxford University Press.

Kristiansen, Tore & Coupland, Nikolas (eds.). 2011. *Standard languages and language standards in a changing Europe* (Standard Language Ideology in Contemporary Europe 1). Oslo: Novus Press.

Labov, William. 1966a. *The Social Stratification of English in New York City*. Washington, DC: Center for Applied Linguistics.

Labov, William. 1966b. The linguistic variable as structural unit. *Washington Linguistics Review* 3. 4–22.

Levshina, Natalia. 2015. *How to do Linguistics with R. Data exploration and statistical analysis*. Amsterdam, Philadelphia: Benjamins.

Levshina, Natalia. 2019. Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology*. De Gruyter Mouton 23(3). 533–572. (doi:10.1515/lingty-2019-0025)

Levshina, Natalia. 2022. Corpus-based typology: applications, challenges and some solutions. *Linguistic Typology*. De Gruyter Mouton 26(1). 129–160. (doi:10.1515/lingty-2020-0118)

Lindau, Mona. 1985. The story of /r/. In Fromkin, Victoria (ed.), *Phonetic linguistics: essays in honor of Peter Ladefoged*, 157–168. Orlando: Academic Press.

Maddieson, Ian. 1988. *Patterns of Sounds*. Cambridge: Cambridge University Press.

Mauri, Caterina & Ballarè, Silvia & Goria, Eugenio & Cerruti, Massimo & Suriano, Francesco. KIParla corpus: A new resource for spoken Italian. In Bernardi, Raffaella & Navigli, Roberto & Semeraro, Giovanni (eds.), *Proceedings of the Sixth Italian Conference on Computational Linguistics*, vol. 2841, 1–7. CEUR Workshop Proceedings.

Milroy, James & Milroy, Lesley. 1997. Varieties and variation. In Coulmas, Florian (ed.), *The Handbook of Sociolinguistics,* 47–64. Oxford: Blackwell.

Paradis, Johanne. 2007. Early bilingual and multilingual acquisition. In Auer, Peter & Wei, Li (eds.), *Handbook of Multilingualism and Multilingual Communication*, 15–44. Mouton de Gruyter. (doi:10.1515/9783110198553.1.15) (Accessed April 30, 2024.)

Parenti, Alessandro. 2001. Sulla semantica dei dimostrativi. *Archivio Glottologico Italiano* 86(2). 174–193.

Pulgram, Ernst. 1964. Structural comparison, diasystems, and dialectology. *Linguistics* 2(4). 66–82. (doi:10.1515/ling.1964.2.4.66)

Renwick, Margaret & Ladd, Robert D. 2016. Phonetic Distinctiveness vs. Lexical Contrastiveness in Non-Robust Phonemic Contrasts. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 7(1). 19. (doi:10.5334/labphon.17)

Sankoff, David. 2004. Variable Rules. In Ammon, Ulrich & Dittmar, Norbert & Mattheier, Klaus J. & Trudgill, Peter (eds.), *Soziolinguistik. Ein internationales Handbuch zur Wissenschaft von Sprache und Gesellschaft*, vol. 2, 1150–1163. Berlin, New York: De Gruyter. (doi:10.1515/9783110171488.2.7.1150)

Schnell, Stefan & Schiborr, Nils. 2022. Crosslinguistic Corpus Studies in Linguistic Typology. *Annual Review of Linguistics* 8. 171–191. (doi:10.1146/annurev-linguistics-031120-104629)

Scobbie, J. M. 2006. (R) as a Variable. In Brown, Keith (ed.), *Encyclopedia of Language & Linguistics (Second Edition)*, 337–344. Oxford: Elsevier. (doi:10.1016/B0-08-044854-2/04711-8)

Siewierska, Anna. 2013. Gender Distinctions in Independent Personal Pronouns. *The World Atlas of Language Structures Online (v2020.3)*. Zenodo. (Ed. Dryer, Matthew S. & Haspelmath, Martin.) (doi:10.5281/zenodo.7385533)

Szmrecsanyi, Benedikt. 2009. Typological parameters of intralingual variability: Grammatical analyticity versus syntheticity in varieties of English. *Language Variation and Change* 21(3). 319–353. (doi:10.1017/S0954394509990123)

Talmy, Leonard. 1985. Lexicalisation patterns: semantic structure in lexical forms. In Shopen, Timothy (ed.), *Language typology and syntactic description*, 36–149. Cambridge: Cambridge University Press.

Talmy, Leonard. 2000. *Toward a Cognitive Semantics*. MIT Press.

Vietti, Alessandro. 2017. Italian in Bozen/Bolzano: the formation of a “new dialect.” In Cerruti, Massimo & Crocco, Claudia & Marzo, Stefania (eds.), *Towards a new standard: theoretical and empirical studies on the restandardization of Italian*, 176–212. Berlin: De Gruyter.

Vietti, Alessandro. 2019. La varietà di lingua come insieme di tratti coerenti: verso una caratterizzazione empirica. *Rivista Italiana di Dialettologia* 43. 11–32.

Vietti, Alessandro & Mereu, Daniela. 2023. Mid vowels at the crossroads between standard and regional Italian. *Sociolinguistica* 37(1). 17–39. (doi:10.1515/soci-2022-0033)

Vietti, Alessandro & Spreafico, Lorenzo. 2016. Lo strano caso di /r/ a Bolzano: problemi di interfaccia. In Iacobini, Claudio & Voghera, Miriam & Savy, Renata (eds.), *Livelli di analisi e fenomeni di interfaccia*, 263–281. Roma: Bulzoni.

Vietti, Alessandro & Spreafico, Lorenzo. 2018. Sprachkontakt in der Phonologie bilingualer Sprecher des Tirolischen. In Rabanus, Stefan (ed.), *Deutsch als Minderheitensprache in Italien. Theorie und Empirie kontaktinduzierten Sprachwandels* (Germanistische Linguistik), vol. 239–240, 49–78. Hildesheim: Georg Olms Verlag.

Villena-Ponsoda, Juan-Andrés & Vida-Castro, Matilde. 2020. Variation, identity and indexicality in southern Spanish: On the emergence of a new variety in urban Andalusia. In Cerruti, Massimo & Tsiplakou, Stavroula (eds.), *Intermediate Language Varieties: Koinai and regional standards in Europe* (Studies in Language Variation), vol. 24, 149–182. Amsterdam: Benjamins. (doi:10.1075/silv.24.07vil)

Wälchli, Bernhard. 2009. Data reduction typology and the bimodal distribution bias. *Linguistic Typology* 13(1). 77–94. (doi:10.1515/LITY.2009.004)

Walker, James A. & Hoffman Michol F. & Meyerhoff, Miriam. 2022. What's in a Lect? Coherence in Phonetic and Grammatical Variation. In Beaman, Karen & Guy, Gregory R. (eds.), *The coherence of linguistic communities: orderly heterogeneity and social meaning*, 71–86. New York: Routledge.

Weinreich, Uriel. 1954. Is a Structural Dialectology Possible? *WORD* 10(2–3). 388–400. (doi:10.1080/00437956.1954.11659535)

Weinreich, Uriel & Labov, William & Herzog, Marvin I. 1968. Empirical Foundations for a Theory of Language Change. In Lehmann, Winfred P. & Malkiel, Yakov (eds.), *Directions for Historical Linguistics. A Symposium*, 97–195. Austin: University of Texas Press.

Wiese, Richard. 2011. The Representation of Rhotics. In Oostendorp, Marc & Ewen, Colin J. & Hume, Elizabeth & Rice, Keren (eds.), *The Blackwell Companion to Phonology*, 1st edn, 1–19. Wiley. (doi:10.1002/9781444335262.wbctp0030)

Wiesinger, Peter. 1990. The Central and Southern Bavarian dialects in Bavaria and Austria. In Russ, Charles (ed.), *The dialect of modern German: a linguistic survey*, 438–519. Stanford: Stanford University Press.

Contacts

alessandro.vietti@unibz.it

massimosimone.cerruti@unito.it

Linguistic Typology at the Crossroads 5-2 (2025): 224-288 ISSN 2785-0943

Governor-driven subjunctive selection: a variationist study from Latin to Romance

SALVIO DIGESTO
CARLETON UNIVERSITY - DEPARTMENT OF FRENCH (OTTAWA, CANADA)

Submitted: 05/06/2024 Revised version: 24/10/2024
Accepted: 07/08/2025 Published: 25/02/2026

Abstract

A key parameter to measure (dis)continuity between Romance languages and the ancestor language, Latin, is mood selection, especially the use of the subjunctive as opposed to the indicative according to syntactic environments and semantic meanings supposedly conveyed. This study explores the trajectory of mood selection in *that*-completive clauses from Latin to modern Romance languages, with a particular focus on Italian. It challenges the assumption that subjunctive selection is semantically motivated, highlighting recent variationist findings that identify the main clause verb's lexical identity as the subjunctive's major predictor. By employing a variationist sociolinguistic approach, this research delineates the evolution of subjunctive selection, revealing that contemporary patterns in Romance languages may reflect a continuation of lexicalization processes initiated in Vulgar Latin, rather than a recent desemanticization phenomenon. This analysis contributes a nuanced understanding of subjunctive selection, offering new perspectives on its function and evolution across Romance languages.

Keywords: subjunctive mood; variationist sociolinguistics; lexicalization; Romance languages; Vulgar Latin; completive clauses.

1. Tracing the conditioning of subjunctive selection

Languages are typically classified according to the presence or absence of a given structural feature, in some cases taking account the feature's frequency or its

distribution. Typologists have examined the subjunctive-indicative distinction in complementation, a well-attested phenomenon in several language families. Romance languages, such as Italian (1), present a dedicated verbal inflectional paradigm inherited from their ancestor language, Latin.[1]

(1) a. *Credo che tutto ritorn**a***$_{[3SG.PRES.IND]}$. (C.511.264)[2]
'I believe that everything comes back.'
b. *Ma può darsi che ritorn**i***$_{[3SG.PRES.SUBJ]}$, *guarda*. (L.16.23)
'But she may come back, look.'

The Italian subjunctive is said to be used to express an attitude of uncertainty and the speaker's attitude towards the truth of the embedded proposition, as opposed to the indicative that is said to convey a more assertive statement (e.g., Whaley 1997; Schachter & Shopen 2007). Subjunctive morphology also supposedly conveys other types of meanings as well, such as indefinite, uncontrolled, or uncertain readings, highlighting its non-assertive nature (Terrell & Hooper 1974; Hooper 1975; Lindschouw 2011). According to Hooper (1975: 123), "complements that are assertions, that is, complements to assertive verbs, have indicative verb forms. […] complements to non-assertive verbs, and (imperative) complements to volitional verbs, all have subjunctive verb forms," suggesting a clear-cut semantic division between subjunctive and indicative contexts. It is said to be a mood fluctuating "between opinion and perception" (author's translation, Binazzi 2015: 259); projecting "a modality of uncertainty and doubt" of the event in question (author's translation, Simone 1993: 80); or representing "the intense and emotive degree, the particular and the personal, the doubt and the unreal, the unexpected and the surprising, the desired and the feared, the extraordinary and the exceptional" (author's translation, Dorigo 1951: 322). In the syntactic literature, the mechanism of mood selection is considered to be "implemented at a distance" (author's translation, Shlonsky 2006: 83), i.e., the main verb determines the type of complementizer and embedded mood, operating by checking mood features (Poletto 2000). This view represents a general assumption in the literature: the governor's semantic

[1] Italian: ita; Indo-European; Romance – Latin: lat; Indo-European; Italic.

[2] The codes identify the corpus, the numeric speaker code, and the line number at which the utterance occurred. Examples from two corpora will be shown throughout this paper, including the C-ORAL-ROM (2005) (C) and the *Lessico di frequenza dell'Italiano Parlato* (1993) (L).

characteristics determine the selection of the 'appropriate' mood for the embedded clause and governors, with *volition*, *emotion*, and *doubt* consistently deemed categorical subjunctive-selecting contexts. This "hereditary feature" that is the Romance subjunctive (author's translation, Cressot 1947: 139) would have inherited from the ancestor language the set of meanings ascribed to it in the Latin era, such as dubitative and hypothetical (Cressot 1947), or uncertain and irrealis (Leone 1949: 13; see also Mellet 1994; Perotti 1996). Italian, in particular, is acknowledged as having "retained from Latin [subjunctive] not only clear formal marking, but also substantial semantic motivation" (Harris & Vincent 1997: 303). However, especially within the Romance language family, whether subjunctive use is triggered primarily by semantic factors remains a matter of some debate in the literature.

1.1. Semantic and pragmatic considerations

While we cannot do full justice to the vast literature on the subjunctive, this section focuses on major trends and influential accounts of its usage and variability, with particular attention to Italian. Key cross-linguistic and foundational studies that have shaped these accounts are also referenced where relevant.

Earlier approaches have typically framed the indicative/subjunctive opposition in terms of a *realis* versus *irrealis* contrast, whereas more recent proposals rely on more refined interpretive parameters (Giorgi & Pianesi 1997: 201). These newer perspectives make the analyst's task more complex, as mood selection becomes closely linked to psychological or attitudinal factors, i.e., the speaker's mental stance at the moment of utterance (e.g., Costantini 2011: 39; Giannakidou & Mari 2015; Portner & Rubinstein 2013). For instance, Wandruszka's analysis of subjunctive mood in completive clauses outlines a hierarchy of subjunctive use based on three semantic groups: volitive, dubitative and factive predicates.

According to Wandruszka (1998: 416), volitive predicates such as those with *volere* 'to want'*, sperare* 'to hope'*, pretendere* 'to pretend', etc. are those that convey the will or intention of the subject. The core semantic property of such predicates is their orientation toward a state of affairs that is not yet realized but is instead desired, prevented, or otherwise projected. As a result, volitive complement clauses do not assert facts; rather, they represent the speaker's stance toward a hypothetical outcome. Wandruszka explicitly notes that a volitive modalized clause is fundamentally distinct from an assertion, and it is precisely this non-assertive nature

that makes the subjunctive particularly salient in such contexts. The subjunctive, in these cases, serves to signal that the embedded proposition is not presented as part of the actual world but as part of the speaker's projected or wished-for reality.

Dubitative predicates (Wandruszka 1998: 418) are characterized as belonging to the broader domain of epistemic modality, i.e., the modality of knowing and believing, where the subject evaluates the truth value or plausibility of a proposition. These predicates range across a continuum, from the speaker's certainty about the non-existence of a state of affairs to varying degrees of uncertainty regarding its existence. What distinguishes this class is not a binary division between true and false, but rather the speaker's degree of commitment to the embedded assertion. While a certain degree of epistemic uncertainty is often sufficient to license the subjunctive, Wandruszka notes that it is not a necessary condition, citing cases such as *non dubito che* 'I don't doubt that' or *sono sicuro che* 'I am sure that', which can still select the subjunctive despite expressing high epistemic commitment. The notion of dubitative or epistemic subjunctive is understood therefore in a broad sense, encompassing predicates like *non credo/credo* 'I don't think/I think', *è possible* 'it is possible', *dubito* 'I doubt', *suppongo* 'I suppose', *pare* 'it seems', and *sono sicuro che* 'I am sure that'. Dubitative or epistemic subjunctive usage is classified across a range of structural types. First, verbs such as *dubitare* 'to doubt' express an absence of belief or knowledge (*non sapere se p* 'not to know if *p*'), and typically select the subjunctive, even under negation (*non dubito che sia intelligente* 'I don't doubt s/he is intelligent'), due to the semantic alignment with conviction rather than factuality. Second, verbs that negate a proposition (*negare che p* 'to deny that *p*') also favour the subjunctive, since they question the truth of the embedded clause. Even when negated themselves (*non negare che p* 'not to deny that *p*'), these predicates retain subjunctive selection. Third, predicates expressing uncertain belief or assumption (*credere* 'to believe', *pensare* 'to think', *ritenere* 'to consider') select the subjunctive when the embedded proposition is not presented as factual. In the presence of negation, interrogative and conditional clauses, this pattern is reinforced (1998: 435-436), as these environments weaken epistemic commitment. Additionally, verbs of communication such as *dire* 'to say' and impersonal constructions such as *dicono che* 'they/people say that', *si dice che* 'it is said that' can trigger subjunctive when they imply reported belief rather than assertion (in a similar vein, Chinellato 2001). Similarly, verbs such as *parere* 'to seem' and *sembrare* 'to seem' tend to select the subjunctive when they convey inference rather than fact (*pare che sia stanco* 'it seems that he is tired'), while modal predicates

like *è possibile che* 'it is possible that', *può darsi che* 'it might be that' nearly categorically require the subjunctive.

Adjectival and nominal predicates indexing epistemic stance also influence mood. Adjectives like *possible* 'possible', *probabile* 'probable', *improbabile* 'improbable' co-occur with the subjunctive, whereas factive adjectives (*certo* 'certain', *evidente* 'evident', *sicuro* 'sure') are more compatible with the indicative. Epistemic nouns such as *dubbio* 'doubt', *opinione* 'opinion', *ipotesi* 'hypothesis', and expressions like *l'idea che* 'the idea that', *l'impressione che* 'the impression/feeling that', favour subjunctive morphology by designating non-factual propositions.

Wandruszka highlights a third major class: factive predicates (which he also refers to as *tematici* or *fattivi di valutazione*; 1998: 418-419). These include evaluative and emotional predicates where the embedded proposition is not in doubt: expressions such as *dispiacersi* 'to regret/to be sorry' or *essere felice* 'to be happy' presuppose the truth of the subordinate clause. What licenses the subjunctive in these cases is not the speaker's uncertainty, but rather the presuppositional status of the embedded content. As Wandruszka observes, one can only experience joy or regret about a state of affairs that is assumed to be real. This presupposition is thus a necessary precondition for the matrix predicate. Moreover, unlike epistemic predicates, factive-evaluative predicates maintain the same truth-value assumptions even when embedded under negation, interrogation, or conditional constructions. The consistent use of the subjunctive in such contexts reflects its core semantic function: to signal that the clause is not part of the speaker's primary communicative act but rather a backgrounded, non-assertive element of discourse.

Within a formal semantic-syntactic framework, Giorgi and Pianesi (1997) aligning in many respects with Wandruszka's discourse-functional approach. Both accounts converge on the idea that subjunctive mood arises when the embedded proposition is not presented as an assertion within the common ground. Subjunctive is consistently licensed by volitional, desiderative, directive, and evaluative predicates, all of which impose modal or counterfactual conditions on the truth of the embedded proposition. In contrast, factives (called *true factives* by Hooper 1975) like *dispiacersi* 'to regret/to be sorry' select the subjunctive due to their emotional-evaluative content and causal structure (Farkas 1992), whereas semifactives such as *sapere* 'to know' select the indicative, given their lack of a modal component. This contrast aligns with Wandruszka's classification of factives and further clarifies why *sapere* 'to know'

resists subjunctive even under negation unless additional semantic or temporal factors intervene, i.e., negation and past tense such as in *non sapevo che* 'I didn't know that'.

In contexts of belief predicates (e.g., *credere* 'to believe', *pensare* 'to think', *ritenere* 'to consider'), Giorgi and Pianesi observe that subjunctive mood is generally licenced. As also noted by Wandruszka (1998), the presence of negative and interrogative operators as well as conditional clauses tend to increase the likelihood of using a subjunctive. Giorgi and Pianesi interpret this pattern in terms of conversational backgrounds and doxastic alternatives: the subjunctive arises when the embedded proposition is not anchored in the speaker's or subject's common ground, but rather evaluated across possible worlds or non-actual scenarios. This conception overlaps with Wandruszka's treatment of epistemic predicates but provides a more formalized grounding. Additionally, they emphasize mood variation under *dire* 'to say' and other *verba dicendi*. While such verbs generally select indicative as the modal base consists of shared facts, and the ordering source is null, operators such as negation or interrogatives can license the subjunctive even under *dire* 'to say', as in *Mario non ha detto che Giuseppe sia impazzito* 'Mario didn't say that Giuseppe has gone mad'. This would demonstrate that mood choice under *verba dicendi* is not purely lexical but sensitive to clause-external conditions.

Although less recent, Schmitt Jensen's (1970) analysis of subjunctive selection attempts to reconcile the competing paradigms in mood selection analysis by returning to structural description while not entirely abandoning semantic considerations. Rejecting the exclusivity of either a purely semantic or purely syntactic model, Schmitt Jensen argues for a dual-layered approach that grounds itself first in observable syntactic structure, i.e., the position and function of elements within the clause, and subsequently seeks interpretive clarity through semantic regularities. Based on a meticulous analysis of a diverse corpus of contemporary Italian, which includes literary texts, essays, journalistic prose, and film dialogues from the 20th century, as well as grammars, Schmitt Jensen identifies four distinct groups of governing predicates (1970: 125-126): a) those that typically govern the subjunctive; b) those that govern the indicative; c) those allowing mood alternation signaling semantic opposition; d) those allowing alternation without any semantic contrast. These types are not derived from theoretical principles but from corpus-driven observation. Only after establishing these distributional categories, he identifies recurrent semantic tendencies within each group, treating these not as defining criteria but as emergent patterns. Within Group A, comprising predicates

that predominantly govern the subjunctive, three main semantic domains recur (1970: 232-233): volition and necessity (*volere* 'to want', *desiderare* 'to desire', *ordinare* 'to command/to order'), doubt and uncertainty (*dubitare* 'to doubt', *temere* 'to fear', *essere incerto* 'to be uncertain'), and subjective or affective evaluation (*essere contento* 'to be happy', *dispiacersi* 'to be sorry/to regret', *è strano che* 'it is strange that'). These predicates typically introduce a non-factual or projected state of affairs and reflect the subject's psychological stance or emotional response rather than a commitment to the truth of the embedded proposition. Group B, by contrast, includes predicates that govern the indicative and presuppose the factuality of the embedded clause (1970: 234). These include epistemically strong verbs such as *sapere* 'to know', *vedere* 'to see', and *essere sicuro* 'to be sure', along with nominal and adjectival expressions like *la certezza* 'the certainty' or *è chiaro che* 'it is clear that', all of which align the embedded content with the speaker's or subject's belief in its reality. Group C consists of predicates that allow both moods, with mood alternation reflecting a semantic contrast (1970: 235-237). For verbs such as *credere* 'to believe', *pensare* 'to think', *dire* 'to say', or expressions like *ho l'impressione che* 'I have the impression/feeling that', the subjunctive marks the embedded proposition as epistemically or pragmatically distanced from the speaker's assertional base, while the indicative aligns it with the common ground. Mood choice in these contexts is said to be often sensitive to discourse factors such as negation, interrogation, or conditional operators, which modulate speaker commitment. Finally, Group D includes cases in which both moods are found without a clearly identifiable semantic opposition (1970: 238). Here, unlike previously discussed accounts, predicates such as *è possibile che* 'it is possible that', *può darsi che* 'it might be that', and impersonal expressions like *si dice che* 'it is said that' exhibit mood alternation without consistent interpretive consequences, raising the possibility that such variation reflects register, regional norms, or stylistic preference rather than semantic contrast. Another interesting aspect raised by Schmitt Jensen's analysis is that the subjunctive functions as a subordination marker, signaling a tighter syntactic and semantic dependency between the matrix and embedded clause. In contrast, the indicative, by virtue of its assertive force, would grant the embedded clause a higher degree of autonomy, effectively reducing the strength of the subordinating link. However, this notion of increased dependency under subjunctive is somewhat problematic, since the author does not provide independent syntactic or semantic diagnostics beyond mood itself to justify such a

hierarchy of integration. This raises the question of whether mood alone can index clause dependency or whether other structural features must be considered.

It is worth noting that across descriptive, formal, and also prescriptive accounts of the subjunctive mood, stylistic and register-based variation in mood choice is acknowledged, though often only marginally. In Wandruszka's analysis, informal, colloquial, or familiar registers are mentioned as contexts in which the indicative may surface in place of the expected subjunctive, even where the semantics would typically require it, for instance, in volitive constructions. Giorgi and Pianesi similarly note intralinguistic variation, such as alternation with *credere* 'to believe', but do not examine register effects, focusing instead on modal semantics within the standard language. Schmitt Jensen, more explicitly, identifies a category of predicates that allow alternation without consistent semantic opposition, and cautiously suggests that stylistic variation, register, and genre may play a role. Yet none of these accounts systematically investigate such alternation beyond intuition or textual observation. These departures from the expected subjunctive are often dismissed as mere matters of style, but the very ease with which they occur raises deeper questions about the force and interpretive necessity of mood marking. If the subjunctive indeed signals epistemic or attitudinal stance, it is unclear why such signaling should be suspended in informal speech. This calls into question theoretical models based primarily on introspective judgments, which may fail to capture the structured nature of vernacular variation. Spontaneous speech data, in fact, can often challenge some widespread assumptions on the conditioning of subjunctive selection (Digesto 2019: 69). For instance, as reported above, a *verbum dicendi* such as *dire* 'to say' may combine with evaluative features, thereby shifting the meaning from that of a purely communicative verb to one that places the proposition within the speaker's opinion or point of view (see Wandruszka 1998; Chinellato 2001, amongst others). This would signal non-engagement with the truth of the embedded proposition (e.g., *dicono* 'they say', *si dice* 'it is said'), as shown in (2):

(2) *Marlon Brando è Marlon Brando. Dicono che* ***sia***$_{[3SG.PRES.SUBJ]}$ *omosessuale.* (C.201.402)
'Marlon Brando is Marlon Brando. People say that he's homosexual.'

However, one might argue that the same indeterminacy can be conveyed when the speaker opts for the indicative counterpart as well (3):

(3) *Qualcuno dice che ci si mett**e**$_{[3SG.PRES.SUBJ]}$ meno.* (C.207.218)
'Some people say that it takes less time.'

Likewise, if we assume that subjunctive morphology signals greater uncertainty, this assumption is challenged by examples such as (4) below. In (4b), the interrogative context suggests that the speaker is uncertain about the truth of the proposition in the complement clause, i.e., which neighbourhood the referent lives in. Nonetheless, the indicative is selected:

(4) a. *Mi sembra che **sia**$_{[3SG.PRES.SUBJ]}$ una felce.* (C.505.141)
'It seems to be that it is a fern.'
b. *Mi sembra che abit**i**$_{[3SG.PRES.IND]}$* a Santa Lucia, *è possibile?* (L.314.28)
'It seems to me that she lives in Santa Lucia, is it possible?'

1.2. Subjunctive mood in quantitative research

Relatively few studies have examined subjunctive selection in Italian through a quantitative lens. Among those that do, two core questions tend to recur: whether the subjunctive is declining, and whether variation in mood choice is semantically governed, particularly by the lexical-semantic class of the matrix predicate. Several studies focus on predicates of volition (Bonomi 1993; Loengarov 2006), hope (Bonomi 1993; Lombardi Vallauri 2003; Santulli 2009), and fear (Bonomi 1993; Santulli 2009), under which the subjunctive is assumed to be categorical, while alternation is expected with predicates of opinion (Bonomi 1993; Lombardi Vallauri 2003; Santulli 2009; Veland 1991). In many cases, the scope is limited to a small set of governors, for example most often *credere* 'to believe', *pensare* 'to think', and *ritenere* 'to consider', which are frequently cited as showing variability and decline. Veland (1991) finds that these verbs of opinion favour the subjunctive, while Bonomi (1993) reports near-categorical selection under opinion (93%) and volition (90%) predicates, with more variability under judgment/evaluative predicates (60%). Gatta (2002) suggests that when the indicative is selected, speakers compensate through other modalizing strategies, such as conditional sentences or expressions like *a mio parere* 'in my opinion', *secondo me* 'to me', and *la mia sensazione* 'my feeling'. Veland (1991: 219) also reports that adverbials encoding [+certainty] can disfavour the subjunctive. Santulli (2009) argues that the subjunctive remains productive in speech, though is

more less likely to occur in first-person contexts (e.g., *penso* 'I think'). Lombardi Vallauri (2003) tests broader semantic classes and finds that the subjunctive is favoured under hope predicates, but disfavoured with verbs of communication.

These studies reveal several interesting patterns, but their divergent data types and methodological choices hinder comparability. While some authors argue for ongoing productivity (Bonomi 1993; Santulli 2009; Veland 1991), others suggest a shift toward the indicative and interpret this as a change in progress (Lombardi Vallauri 2003; Schneider 1999). Yet, none of these claims is supported by diachronic evidence. Voghera (1993) points to low overall rates in speech, which she interprets as part of a broader decline of the subjunctive. Overall rates vary widely, from 2% (Voghera 2001) to over 80% (Santulli 2009), but are based on different corpora and non-uniform criteria, precluding direct comparison.

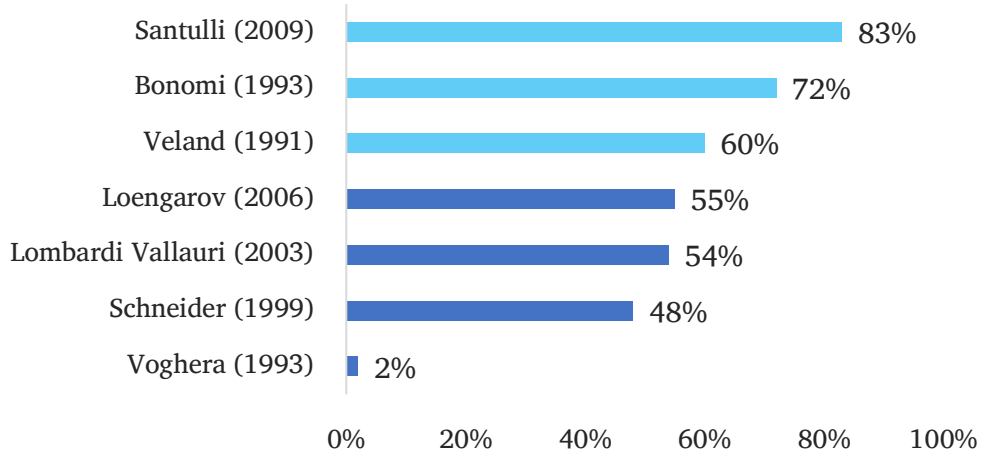


Figure 1: Comparison of the rate of subjunctive selection in previous quantitative research. With the exception of Voghera (1993), overall rates were calculated on the basis of the results provided by the authors in their publications. Results of the studies that made use of LIP corpus are in dark blue.

In terms of testing semantic conditioning and the contribution of some lexical types, some studies impose some restrictions by extracting only embedded forms of *essere* and *avere* (Lombardi Vallauri 2003; Loengarov 2006) or restrict subject person to either first (Santulli 2009) or third (Loengarov 2006). These a priori constraints may obscure the full range of variation, violating the Principle of Accountability (Labov

1972), since the subjunctive might occur with other governors and contexts in the selected data[3].

The data sources are similarly heterogeneous. In addition to spontaneous speech corpora, studies have drawn on newspapers (Bonomi 1993; Santulli 2009; Veland 1991), literary texts (Soliman 2002), text messages (Santulli 2009), online forums (Loengarov 2006; Santulli 2009), and television talk shows (Gatta 2002). Even studies using the same dataset, such as the LIP corpus, differ in their selection criteria, excluded contexts, and analytical procedures, whether they report raw counts or rates. As a result, a single governor may be found to favour the subjunctive in one study and disfavour it in another. For example, *credere* 'to believe' is reported as favouring in Veland (1991) and disfavouring in Bonomi (1993); *ritenere* 'to consider' appears to favour in Santulli (2009) and disfavour in Schneider (1999).

1.3. Towards an account of the lexicalization of the subjunctive in Romance

Following Poplack's seminal work on subjunctive variability (Poplack 1990; 1992; also Poplack et al. 2013), recent variationist sociolinguistic research has addressed the above-mentioned issues and examined the use of the subjunctive in running discourse, particularly in complement clauses, highlighting a trend that is shared across the varieties of Romance languages examined (e.g., Italian, Canadian French, Parisian French, Brazilian Portuguese, Mexican Spanish) and uncovering a systematic selection process guided more by lexical governors than semantic roles (Digesto 2021; Kastronic 2016; Poplack et al. 2018). This pattern suggests a path of development divergent from prescriptive norms and from common assumptions that the subjunctive is used in discourse in contexts that convey meanings such as uncertainty, doubt, volition, necessity, etc., revealing instead a trajectory towards *lexical routinization*—a process where the selection of the subjunctive mood becomes progressively restricted to a select array of lexical triggers (Poplack et al. 2018). This lexicalization in discourse challenges the traditional understanding of the subjunctive

[3] The reader may note that, compared to earlier quantitative studies using the LIP corpus, the number of verbs analyzed in the current study differs. This is due both to the fact that previous studies imposed *a priori* lexical choices, whereas the present analysis considered all governors that triggered the subjunctive in context (i.e., determined corpus-internally), and to the restriction of the current study to a sub-section of LIP containing only spontaneous, unscripted conversations. For a detailed discussion of stylistic and social variation in contemporary speech, see Digesto (2019, 2021).

as semantically driven, positing an alternative account where syntactic environments and lexical collocations underpin its usage. The insights gleaned from examining speech data from a several Romance languages and a wide temporal range have been pivotal in characterizing the subjunctive not as a feature constrained by semantic properties but as one deeply embedded within the lexicon of certain governing verbs. Poplack and her associates (1992; 2013) notably illustrate this phenomenon in French, where the verbs *falloir* 'it is necessary', *vouloir* 'to want', and *aimer* 'to like' significantly trigger the selection of the subjunctive in speech. This phenomenon, found also in a study of subjunctive usage in Parisian French (Kastronic 2016), indicates a systematic entrenchment of the subjunctive within particular lexical environments cross-varieties. Likewise, contrary to prevailing assumptions in the literature regarding the semantic/pragmatic motivation for selecting the subjunctive mood in embedded completive clauses in Italian, Digesto (2019; 2021) showed that the use of subjunctive is mainly restricted to a handful of main clause verbs (e.g., *bisognare* 'it is necessary'*, credere* 'to believe'*, pensare* 'to think'*, sembrare* 'it seems') and a single embedded verb (suppletive forms of *essere* 'to be'). Similar trend, i.e. a less productive use of the subjunctive in speech that is mainly triggered by a handful of matrix verbs and embedded verbs, particularly favoured with irregular/suppletive morphology, is observed in other Romance languages examined through the lens of variationist method, such as Portuguese and Spanish (Poplack et al. 2018; also Berlinck 2019). This pattern towards lexicalization, identified by some scholars as a case of subjunctive attrition (Bybee et al. 1994), would mark a departure from a semantically-based usage of the subjunctive through desemanticization or 'semantic bleaching'. In other words, the loss of a semantic contribution, through a process of *ritualization* "brought about trough routine repetition" (Haiman 1994: 3; see also Bybee 2003), results in *obligatorification* (Lehmann 1995: 12), or what is referred to elsewhere, in variationist research, as *lexical routinization* (Poplack et al. 2018: 238).

Lexical routinization implies that semantics contribute minimally to none to the choice of subjunctive, favouring structural and lexical constraints over semantic underpinnings (Haiman 1994; Bybee 2003; Lehmann 1995). While variationist research on contemporary usage of the subjunctive mood, based on recordings of actual speech data highlighted common patterns in Romance pointing mostly to a lexicalized use in discourse, variationist sociohistorical research showed that similar trends have been operative in Italian since at least the 16th century (Digesto 2019) and since the 19th century in Canadian French (Poplack et al. 2013) and in Brazilian Portuguese (Berlinck 2019), at our knowledge, the oldest historical benchmarks investigated by following

variationist methodology. Results showed that, contrary any expectation of a semantically-based use of the subjunctive in previous stages of these languages, the grammar of usage has remained remarkably stable, with no semantic contributions evidenced in previous stages of the targeted varieties but rather a persistence of a lexicalized pattern as outlined above.

We are still left with the question: Can the lexical conditioning of the subjunctive be considered an innovative pattern that emerged in the vernacular daughter languages, assuming that in the ancestral language, Latin, the subjunctive mood was triggered by semantic factors? Or is it a case of inheritance and transmission from Latin, suggesting that the lack of semantic contribution *already* characterized the use of the subjunctive at its roots? Building on previous variationist research, this contribution explores the use of the subjunctive within a corpus of Latin data, in a variationist framework. This approach ensures comparability with results observed in Italian, as well as more broadly across other Romance languages, particularly French and Portuguese.

2. The evolution of the subjunctive from Latin to Romance

It is generally observed that a distinction between *meaningful* and *meaningless* morphology was found in Latin, which stems from the observation that the subjunctive mood was used in various syntactic environments, in both main and various subordinate clauses. While its application in main clauses is considerably more extensive in Latin than in modern Romance languages (Harris 1978), the subjunctive is predominantly associated with *subordination* (Magni 2009: 244; Noonan 2007; Jespersen 1924; Palmer 2001). Traditionally, the opposition between indicative and subjunctive moods in Latin is seen as carrier of semantic differences (Harris 1974: 169). However, during the Vulgar Latin period, the subjunctive would have begun to function more as a conventional or stylistic marker in certain contexts, indicating a shift towards a predominantly formal and mechanical attraction of the embedded mood (Handford 1947: 149). It is commonly accepted that the shift from paratactic to hypotactic syntax prompted a reorganization of independent clauses into more complex, dependent structures (Haudry 1973; Harris 1978; Murphy 2008). This reorganization is said to mirror the subjunctive's original optative value in earlier independent structures, rendering its morphology *meaningful*, especially when influenced by verbs expressing volition or order, as well as by verbs of hindering or preventing (Magni 2009: 245; Bybee et al. 1994).

PARATAXIS	>	HYPOTAXIS
volo; veniat$_{SBJV}$		*volo ut veniat*$_{SBJV}$
'I wish it; let him come'		'I want him to come'

Figure 2: Shift from parataxis (with subjunctive used in the main clause) to hypotaxis (with subjunctive used in the subordinate clause) in Latin. Adapted from Harris (1978: 168).

The shift to hypotactic structures led to the gradual permeation of the subjunctive in syntactic contexts traditionally associated with the indicative, such as indirect questions and adverbial sentences, which are syntactically considered *non-harmonic* with the original main clause meaning (Magni 2009: 247-250). This broadened use of the subjunctive denotes a departure from its original meaning and rather becoming a marker of subordination (Magni 2009: 260; Harris 1978).

Another significant syntactic development affecting subjunctive use in subordinate clauses, and its inheritance in Romance languages is the typological shift from OV (Object-Verb) to VO (Verb-Object) word order (Murphy 2008):

- *Left-branching subordination* included a) infinitival constructions, b) participial constructions, and c) subjunctive mood with no additional marker of subordination;
- *Right-branching subordination* included overt pre-posed subordinating conjunctions, i.e. subordination is marked pre-verbally.[4]

This shift led to an increase in explicit subordinating conjunctions and reinforced the subjunctive's syntactic subordinating function, moving away from a semantically-driven selection process. It suggests a preference for finite complementation over non-finite forms such as the *Accusativus cum Infinitivo*, promoting a widespread use of the subjunctive in right-branching structures. Despite some claims of SVO being the colloquial norm in earlier Latin (Pinkster 1990: 188), this shift likely furthered the use of the subjunctive as a standard marker of subordination in Latin, weakening any hypothesis of a semantically-based usage.

Similarities and discrepancies related to the type of matrix verbs that select the subjunctive are well-documented between Latin and Italian (and Romance in general): where in Latin verbs of thinking and saying typically select the indicative,

[4] Herman (1989) similarly noted, providing quantitative evidence, that in Late Latin, there was a syntactic competition between finite complements and non-finite complements (*Accusativus cum Infinitivo*). He pointed out that "before the verb, AcI is practically the only possibility, whereas after the verb, the binary choice between AcI and *quod/quia* clauses becomes possible" (*ibid.* 133).

supposedly categorically, Italian sees permeation of the subjunctive into these contexts. On the other hand, necessity, emotive and volitive verbs do not contrast in terms of mood selection between the ancestor and the daughter language—all supposedly triggering subjunctive categorically (Harris & Vincent 1997: 67; Disterheft & Viti 2010: 248; Clackson 2011: 137; Magni 2009; Palmer 2001). However, the main discrepancy lies in the type of constructions these main-clause verbs govern in the targeted languages: verbs of saying and thinking, as well as volitive verbs are generally considered quasi-categorical contexts for selecting *non-finite* constructions in Latin, as opposed to Italian, and to some extent in other Romance languages.

The expected outcome in relation to the type of main predicate can be schematized in the following way:

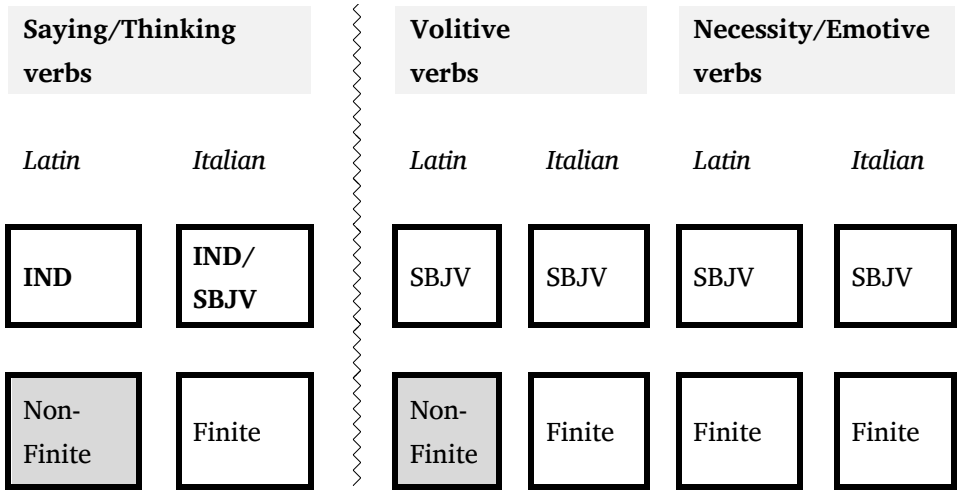


Figure 3: Prediction of mood selection according to the type of main predicate, in Latin and Italian, based on what is generally reported in the literature.

3. A variationist approach on subjunctive selection

3.1. Methodology

The approach to understanding subjunctive selection through the lens of variationist sociolinguistics offers a methodologically robust framework, placing empirical analysis at its core. This perspective posits that language variation is not random but is systematically structured (*ordered heterogeneity*; Labov 2001). It pivots on the fundamental notion that speakers, embedded within their communities, make use of alternate linguistic forms to express identical meanings or functions. This is the

abstract construct of the *linguistic variable* (Labov 1972; Labov 1984). Within this framework, the subjunctive mood is analyzed as a variable element, a variant form that alternates in discourse with other forms (such as the indicative) to express the same meaning or function in discourse (Poplack 2011: 212; Poplack & Levey 2010: 398). However, for many accounts of mood selection, the subjunctive and the indicative are ascribed distinct functions, which posit a methodological caveat when considering these two moods as variants that alternate to express equivalent meanings or functions in discourse. One way to circumvent this issue could be via the identification of those contexts where the subjunctive is *supposed* to occur and subsequently ascertain whether it actually does occur there in the data. However, previous research has highlighted the lack of agreement on which contexts or meanings *exactly* should trigger the subjunctive and therefore the rules governing subjunctive usage remain somewhat generally elusive (see Poplack et al. 2013, for French; Digesto 2019, for Italian). Following Poplack (1992, 2013, and Poplack at al. 2018), we identify the contexts of use *corpus-internally*, by locating all the contexts where the subjunctive is *actually used* to determine where it *could* be used. We identified all the unambiguous subjunctive forms and exhaustively extracted them from the dataset. This method enables the identification of which main clause verbs, called *governors*, were accompanied by the subjunctive, yielding the list of all the governors that selected the subjunctive in the dataset. Once the lexical identities of the governors are identified, the next step is to go back through the data and to extract all of the variants that compete with the subjunctive. This process is known as *circumscribing the variable context*. It allows to fully account for the variation in discourse by including "every case where the variable element occurs in the relevant environments as we have defined them" (Labov 1972: 72), but also where it could have occurred but it *did not* (Poplack & Tagliamonte 2001: 89). Following Poplack et al. (2019: 229), the variable context is here defined as every tensed embedded clause governed by a matrix element—whether verbal, nominal, or adjectival—that triggered the subjunctive at least once.

During the extraction phrase of the data, homophony between the subjunctive and the indicative was considered, in order to select unambiguous subjunctive morphology. Homophony is found in Latin for a wide variety of forms, mainly belonging to the active voice and overlapping between future indicative and perfect subjunctive. For instance, 1st conjugation verbs (e.g., *amo* 'I love') between future perfect indicative and perfect subjunctive: 2SG (*amaveris*), 3SG (*amaverit*), 1PL

(*amaverimus*); 2PL (*amaveritis*), 3PL (*amaverint*); 2nd conjugation verbs (e.g., *moneo* 'I warn'): 2PL (*monueritis*) and 3PL (*monuerint*); 3rd (e.g., *dico* 'I say') and 4th conjugation verbs (e.g., *audio* 'I hear'): 1SG future indicative and present subjunctive (*ducam*) at the active voice as well as at the passive voice (*ducar*); a few forms of the future perfect indicative and perfect subjunctive such as 2SG (*duxeris*), 3SG (*duxerit*), 1PL (*duxerimus*), 2PL (*duxeritis*), 3PL (*duxerint*); as well as 2SG, 3SG, 1PL, 2PL, 3PL between perfect future indicative and perfect subjunctive of almost all irregular verbs such as *sum*, *possum*, *eo*, *volo*, *nolo*, *malo*, *fero* and *do*.[5] On the other hand, homophony in Italian is limited to 2nd person singular of the present indicative and the present subjunctive with first group verbs (*-are*), 1st person plural of the present indicative and the present subjunctive of all three conjugation groups, and finally, 2nd plural of the simple past indicative and the imperfect subjunctive of all three conjugation groups.

The method applied in the current study allows for direct comparability between findings from synchronic and diachronic research on Italian and other Romance languages, and those derived from Latin data. We implemented the *Comparative Variationist Method* (Poplack & Tagliamonte 2001): rather than modeling change as proceeding directly from stage A to B, thereby overlooking the transition period from point A to point B in time and consequently portraying change as very abrupt, characterising the comparative method (Campbell 2013), we assume that a given language inherits the variants of a particular function from its ancestor, along with the linguistic conditioning of their variability. Thus, the genetic relationship is determined on the basis of the *conditioning* on variability, as revealed by which factors trigger the use of the subjunctive in discourse, rather than relying solely on the forms themselves.

3.2. Data

The contemporary data analyzed in this study come from two well-established corpora of spoken Italian: the *Lessico di frequenza dell'italiano parlato* (LIP; De Mauro et al. 1993) and the *C-ORAL-ROM Corpus of spoken Romance languages* (Cresti & Moneglia 2005). Both corpora include recordings of conversations, dialogues, and monologues, primarily collected in major urban centres such as Milan, Florence,

[5] 'I am', 'I can', 'I go', 'I want/I wish', 'I do not want/I do not wish', 'I prefer', 'I carry/I bring', 'I give'.

Rome, and Naples. For the purposes of this analysis, only those recordings capturing naturally occurring, unscripted speech in informal contexts were retained, while any data from institutional or pre-planned settings (e.g., televised interviews or press briefings) were systematically excluded. While invaluable for empirical research, these corpora are not without limitations. As previously noted (Digesto 2019: 44; Digesto 2021), there are sampling biases: Florentine data dominate the informal speech in C-ORAL, and LIP lacks socio-demographic metadata, making it difficult to verify whether speakers are truly representative of their local communities; where speaker information is available, the data tend to skew toward individuals with high levels of education. Nonetheless, these corpora offer a rare opportunity to conduct quantitative analysis of subjunctive usage in *spontaneous speech*. All data were manually concordanced, and all unambiguous instances of subjunctive morphology were retained for analysis. Given the prestige associated with the subjunctive, its frequency may be somewhat inflated in these sources. However, as shown in previous studies (e.g., Poplack & Tagliamonte 2001; Poplack & Levey 2010), while frequency may vary with social factors, the structural conditioning of variation is typically stable across genres and speaker groups.

3.3. Selecting the Latin benchmark

Variationist research mainly focuses on spoken usage, and particularly *vernacular*, i.e. "the style in which the minimum attention is given to the monitoring of speech" (Labov 1972: 208), since this is not only characterized by inherent variability but it also appears to be "the most systematic data for linguistic analysis" (Labov 1984: 29). This might be particularly challenging for analyzing early language stages or ancestral languages, as significant changes often emerge in spoken rather than written form. Without recordings or direct evidence of these stages, older texts cannot be reliably considered accurate representations of vernacular speech (Ayres-Bennett 2000; Nevalainen & Raumolin-Brunberg 2017: 26; Romaine 1982: 14). Written texts often fall short of capturing non-standard features; even texts intended to reflect spontaneous speech are shaped by prescriptive norms or linguistic ideology. The inclusion of non-standard linguistic features or sociolinguistic stereotypes in texts might also reflect the author's intent to depict lower-class/uneducated speech rather than an authentic portrayal of vernacular norms (Labov 1994: 11). To address this,

researchers have turned to speech-like surrogates, notably popular theatrical plays that aim to replicate the daily speech of different social classes, including the lower, working, and upper/aristocratic classes (Palmer 2001; Haudry 1973; Murphy 2008).

Since this research aims to explore the continuity and innovation of subjunctive use from Latin to Romance, with a specific emphasis on Italian, to assess how and to what extent contemporary usage aligns with historical practices in the ancestor language, it is essential to 1) select the specific variety of Latin for investigation and 2) choose texts that resemble speech. First, this emphasizes the need to focus on Vulgar Latin (VL), the spoken precursor to the Romance languages, rather than Classical Latin, primarily associated with written texts. Second, the choice of the data to examine fell onto an example of material that reflects features of VL and non-standard linguistic behaviour: we selected the *Satyricon*, a comic romance attributed to Petronius and dated to the 1st century (Durante 1981: 29). Specifically, the chapters of the *Cena Trimalchionis* ('Trimalchio's dinner party') constitute the VL corpus adopted in this study. Despite the limitation of the sample, with 12,576 words, it offers a unique glimpse into the informal, speech-like Latin that deviates from the normative Classical tradition.

This distinctive source allows a deep dive into the vernacular Latin of a different social stratum, showcasing a vivid portrayal of 'popular' speech through the dialogue of Trimalchio and his rude dining companions. This representation is celebrated as "the happiest achievement of ancient realism" (Scarsi 1996: xix), filled with elements like stories, proverbs, stereotypes, gossip and beliefs and providing more naturalistic forms of expression that are valued in variationist research. Though not a modern theatrical work, the *Satyricon* was intended for live performance to an audience of average citizens (Scarsi 1996), distinct from the elite circles often depicted in classical texts. Through Petronius's work, we gain insights into the *sermo plebeius*, the common speech of the first century A.D. (Palmer 1988: 153), offering a valuable perspective for understanding the use of the subjunctive in Vulgar Latin and its evolution in Romance. The chapters of the *Cena* not only contain a vocabulary that is "forceful, coarse, often indecent" (Palmer 1988: 152-153), but they also lay bare some explicit linguistic choices diverging from the standard of Classical Latin. For instance, errors of declension (7), vernacular verb forms (8), as well as the non-standard use of indicative instead of the normative subjunctive (9).

(7) *Lacte gallinaceum si quaesieris invenies.*
[*lacte* instead of the normative accusative form *lac*] (XXXVIII.249)[6]
'You can have cock's milk if you desire.'

(8) *Itaque statim urceatim plovebat*$_{[IMPRF.IND.3SG/ACT]}$.
[instead of *pluebat*$_{[IMPRF.IND.3SG/ACT]}$] (XXXVIII.249)
'And immediately the rain came down in bucketfuls.'

(9) *Cum interim nemo curat quid annona mordet*$_{[PRES.IND.3SG/ACT]}$.
[instead of *mordeat*$_{[PRES.SBJV.3SG/ACT]}$] (XXXVIII.249)
'Meanwhile, nobody cares how high the market price is.'

Despite the limitations of the historical dataset, we can use this data to uncover the internal conditioning of the subjunctive, which serves as major component of comparative analysis. Through systematic comparison of VL and Italian, and subsequently Romance, we will be in a position to ascertain whether and to what extent the subjunctive usage differs from the ancestor language to a daughter language.

3.4. Operationalizing hypotheses

Building on previous variationist studies (Poplack et al. 2013; Poplack et al. 2018), this study investigates semantic and lexical conditioning of subjunctive selection in Italian and Latin, focusing on the following parameters:

i. The semantic category of the governor
ii. The lexical identity of the governor
iii. The structure of the matrix clause and polarity
iv. The presence of other indicators of non-factual modality
v. The type of complementation

As outlined above, a recurrent claim in the literature is that specific semantic properties of governing verbs drive subjunctive selection. This study tests such claim by coding each occurrence according to parameters (i), (iii), and (iv), as detailed in

[6] The codes identify the chapter of the Satyricon and the line number at which the utterance occurred.

the ensuing section. Additionally, every token was coded based on the lexical identity of the governor. This analysis focused on three aspects: 1) the frequency of subjunctive usage for each governor, 2) the proportion that each governor represents within the entire governor pool, and 3) the extent to which each governor accounts for subjunctive morphology. A high score on these measures could indicate limited productivity and suggest a trend toward lexicalization of the subjunctive. If the lexical identity of the governor contributes to subjunctive selection *independently* of the intended meaning, this could signify that the productivity of the subjunctive is declining, and its use is confined to specific lexical governors.

Going beyond the envelope of variation, we also explore the type of complementation preferred by each governor, aiming to discern the extent to which the subjunctive use in Latin was influenced by the type of the complement (finite versus non-finite). This analysis could reveal the degree of syntactic contrast and to what extent this affects mood selection, particularly subjunctive, in Latin as opposed to the daughter languages.

In variationist research, a quantitative approach to modeling variation and change is paramount:

> The effect of a given factor is inferred by comparing its individual rate of subjunctive to the overall rate for the pooled data […] If a factor shows a rate of subjunctive selection higher than the overall rate, its effect is deemed favouring. The greater the difference, the stronger the effect. Likewise, the same method is applied when deeming the effect disfavouring (rate of subjunctive selection of a given factor lower than the overall rate). A neutral effect is deemed when there is no substantial difference between the rate of subjunctive of a given factor and the overall rate. (Digesto 2021: 13).

Employing the same analytical method and quantitative framework as earlier variationist research enables cross-linguistic comparison of findings. By examining the aforementioned parameters, both in isolation and in combination, this study seeks to examine whether the subjunctive was a productive and semantically-driven mood in Latin, and whether its productivity and semantic contribution were subsequently weakened in Romance languages. This could be evidenced by the diminishing role of semantic factors in the selection of the subjunctive and its increasing entrenchment in specific lexical environments (Lehmann 1995; Bybee 2003; Haiman 1994), as

demonstrated by the lexical routinization of the subjunctive in Romance (Poplack et al. 2018).

4. Results: contemporary spoken Italian

Overall results are based on the extraction and coding of the dependent variable, i.e., the grammatical mood (subjunctive or indicative). A first observation is that mood selection under governors that triggered the subjunctive at least once in the dataset shows robust variability: the subjunctive is the majority variant, occurring in 66% of cases (N = 404/616), followed by the indicative at 34% (N = 212/616), confirming that alternation remains frequent and that subjunctive mood is apparently quite productive in discourse. This also establishes our benchmark for interpreting any favouring effects of independent factors, particularly semantics.

4.1. The semantics of the governing verb

Among the core claims in the literature on subjunctive selection is that mood choice is semantically conditioned, determined in part by the meaning of the matrix predicate. To empirically test this claim, this study follows the variationist methodology developed by Poplack (1992; 2013), which evaluates such hypotheses against actual usage patterns in discourse rather than analyst assumptions. Operationalizing semantic claims is a crucial but often difficult task for the analyst. This is due on the one hand to the lack of consensus across studies on exactly which semantic categories, meanings, and governors should be considered for a study of the subjunctive in speech, and on the other hand, to the fact that many of the meanings proposed in previous research stem from analyst introspection into the supposed beliefs, feelings, or desires of the speaker. This is problematic because, as Poplack (2013: 162) points out, there is often no objective way to reconstruct speaker intent or underlying meaning from discourse, let alone to operationalize and test it. In the present study, I follow the method implemented by Poplack (1992, 2013), which circumvents this issue to some extent by relying on observable patterns in naturalistic data to test the contribution of semantic factors to subjunctive selection.

Rather than reproducing previous classifications that often rely on traditional notions of *volition*, *emotion*, or *opinion*, this study adopts a three-way typology of

predicate classes based on Wandruszka's (1998) descriptive framework: *volitive* (10a), *dubitative* (10b), and *factive* (10c).

(10) a. VOLITIVE

Ma po' mi comprava i' pesce. Voleva che lo ***cucinassi***$_{[3SG.IMPRF.SUBJ]}$. (C.001.398)

'But then s/he would buy me fish. S/he wanted me to cook it.'

b. DUBITATIVE

Mi sembra che ***sia***$_{[3SG.PRES.IND]}$ *una felce.* (L.120.81)

'It seems to be that it is a fern.'

c. FACTIVE

Non *mi dispiace che tu* ***faccia***$_{[3SG.PRES.SUBJ]}$ *con me questo compito.* (L.10.21)

'I don't mind you doing this homework with me.'

If semantic conditioning underlies subjunctive selection, we expect consistent patterns across members (i.e., governors) of each class: the subjunctive should appear categorically with volitive predicates, variably with dubitative ones, and be disfavoured with thematic-factives. Crucially, no single lexical item should determine the behaviour of an entire category; rather, the consistency of effect across predicates within each class is the relevant diagnostic. This approach allows for a more systematic test of semantic hypotheses and avoids the pitfalls of relying on a handful of a priori selected governors, which has been a common limitation in previous research.

The results for this factor group, as presented in Table 1, show clear differences in subjunctive rates across the three categories.

Semantic Class	% Subj	N
Factive	95%	21/22
Volitive	82%	108/131
Dubitative	60%	272/450
Communicative	23%	3/13
Total	**66%**	**404/616**

Table 1: Subjunctive selection rates by semantic class of the governing verb.

A first important result is that all semantic classes exhibit variation in mood selection, albeit to differing degrees, including the volitive class (82%), "widely touted as the

wellspring of subjunctive morphology" (Poplack et al. 2018: 234). Contrary to expectations, the strongest favouring effect emerges in the factive class (95%), followed by volitive predicates (82%), while dubitative governors slightly disfavour the subjunctive (60%). A small number of tokens headed by *dire* 'to say' pattern rather with *verba dicendi*, aligning more with assertive than epistemic usage, and show a markedly low subjunctive rate (23%), as in (11).

(11) a. *Io, in terza media, c'avevo una che diceva che Leopardi* ***era***$_{[3SG.IMPRF.IND]}$ *immortale.* (C.087.55)
'Me, when I was in middle school, I had a teacher who used to say that Leopardi was immortal.'

b. *Non si può dire che nel terzo mondo* ***siano***$_{[3SG.PRES.SUBJ]}$ *tutti cattivi. (*L.412.140)
'One cannot say that in the Third World they are all bad.'

It is worth noting that volitive and dubitative classes account for the bulk of the dataset (94%, n = 581/616) and of the subjunctive forms produced in discourse (94%, n = 380/404). On the surface, one could claim that this hierarchy between volitive and dubitative governors supports Wandruszka's as well as other accounts of subjunctive selection (Loengarov 2006; Giorgi & Pianesi 1997; Schmitt Jensen 1970, among others) placing variability primarily within the epistemic domain of opinion or dubitative predicates. However, this presupposes internal uniformity: that every member of a given semantic class would pattern alike. Closer examination reveals that this is not the case. Variation persists within each class, undermining the assumption that semantic category alone can predict mood selection consistently.

Volitive Governor	Translation	% Subj	N	% Data	% Subj Morph
bisognare	to be necessary	93%	41/44	34%	38%
(non) volere	(not) to want	84%	16/19	15%	15%
sperare	to hope	93%	14/15	11%	13%
bastare	to be enough / suffice	50%	6/12	9%	6%
aspettare	to wait / to expect	50%	4/8	6%	4%
è importante	it is important	100%	4/4	3%	4%
fare sì	to ensure / to make sure	75%	3/4	3%	3%
esigere	to demand	100%	3/3	2%	3%

Volitive Governor	*Translation*	*% Subj*	*N*	*% Data*	*% Subj Morph*
avere paura	to be afraid	75%	3/4	3%	3%
è inutile	it is useless	50%	2/4	3%	2%
richiedere	to require / to demand	67%	2/3	2%	2%
non avere senso	to make no sense	100%	1/1	1%	1%
evitare	to avoid	100%	1/1	1%	1%
non preoccuparsi	to not worry	100%	1/1	1%	1%
preferire	to prefer	100%	1/1	1%	1%
sentire il timore	to feel fear	100%	1/1	1%	1%
l'importante è	the important thing is	100%	1/1	1%	1%
servire	to be needed / to be of use	100%	1/1	1%	1%
lasciare	to let / to allow	100%	1/1	1%	1%
fare in maniera	to ensure	50%	1/2	2%	1%
fare in modo	to ensure / to make sure	100%	1/1	1%	1%
Total		**82%**	**108/131**		

Table 2: Subjunctive selection across volitive predicates.

As shown in Table 2, while the overall subjunctive rate across volitive governors is 82%, closer inspection reveals considerable variability among individual verbs. Some governors strongly favour the subjunctive, such as *bisognare* 'it is necessary' and *sperare* 'to hope' (93%), while others favour it to a lesser degree, as with *volere* 'to want' (84%). A few select the subjunctive categorically (*è importante* 'it is important', *evitare* 'to avoid', *preferire* 'to prefer', etc., all at 100%), while others disfavour it or yield markedly lower rates, including *fare sì* 'to ensure' (75%) and *bastare* 'to be enough' (50%). Strikingly, just three verbs, *bisognare* 'it is necessary', *volere* 'to want', and *sperare* 'to hope', account for 60% of all volitive tokens (78/131) and 66% of the subjunctive morphology observed (71/108), with *bisognare* 'it is necessary' alone representing 34% of the data and 28% of the subjunctive forms. These patterns mirror previous variationist results (Poplack et al. 2018; Digesto 2021), underscoring the extent to which a small number of predicates can drive the overall favouring effect.

A similar trend is observed within the class of dubitative governors, which comprises the bulk of the dataset: 73% of all tokens and 67% of the total subjunctive morphology.

Dubitative Governor	Translation	% Subj	N	% Data	% Subj Morph
(non) pensare	to (not) think	73%	64/88	20%	24%
(non) sembrare	to (not) seem	75%	48/64	14%	18%
(non) credere	to (not) believe	79%	45/57	13%	17%
non è	to not be	33%	37/112	25%	14%
(non) parere	to (not) appear	58%	15/26	6%	6%
può darsi	to be possible / it might be	58%	11/19	4%	4%
(non) essere sicuro	to (not) be sure	63%	5/8	2%	2%
(non) dire	to (not) say	21%	3/14	3%	1%
avere la sensazione	to have the feeling	100%	3/3	1%	1%
non è detto	to not be certain	100%	3/3	1%	1%
ritenere	to consider	100%	3/3	1%	1%
supporre	to suppose	100%	3/3	1%	1%
avere l'impressione	to have the impression	50%	2/4	1%	1%
ci sta	to be possible	100%	2/2	0%	1%
è impossibile	to be impossible	100%	2/2	0%	1%
mettere	to assume	100%	2/2	0%	1%
non sapere	to not know	33%	2/6	1%	1%
non succedere	to not happen	100%	2/2	0%	1%
presupporre	to presuppose	67%	2/3	1%	1%
può essere	to be possible	100%	2/2	0%	1%
(non) è possibile	to be (not) possible	33%	1/3	1%	0%
assicurarsi	to make sure	100%	1/1	0%	0%
avere il dubbio	to have doubt	100%	1/1	0%	0%
calcolare	to calculate	50%	1/2	0%	0%
controllare	to check	100%	1/1	0%	0%
dedurre	to deduce	100%	1/1	0%	0%
dubitare	to doubt	50%	1/2	0%	0%
è ovvio	to be obvious	33%	1/3	1%	0%
immaginare	to imagine	25%	1/4	1%	0%
non avere il dubbio	to have no doubt	100%	1/1	0%	0%
non è vero	to not be true	50%	1/2	0%	0%
presumere	to presume	100%	1/1	0%	0%
reputare	to deem / to consider	100%	1/1	0%	0%
rischiare	to risk	100%	1/1	0%	0%
trovare	to find	50%	1/2	0%	0%
verificare	to verify	100%	1/1	0%	0%
Total		**60%**	**272/450**		

Table 3: Subjunctive selection across dubitative predicates.

Despite the overall rate of subjunctive selection being 60% (lower than the corpus-wide average of 66%), not all verbs in this category conform to a disfavouring pattern. Some governors, such as *avere la sensazione* 'to have the feeling', *supporre* 'to suppose', and *ritenere* 'to consider', show categorical selection of the subjunctive (100%), while others, such as *pensare* 'to think' (73%), *sembrare* 'it seems' (75%), and *credere* 'to believe' (79%), exhibit a clear favouring effect. Verbs like *(non) essere sicuro* '(not) to be sure' (63%) hover close to the class average, suggesting a slightly favouring effect. By contrast, others such as *parere* 'it seems' (58%) and *può darsi* 'it might be' (58%) show slight disfavour, and *non è* 'it is not' (33%) and *dire* 'to say' (21%) markedly disfavour the subjunctive. Notably, three verbs, *pensare* 'to think', *sembrare* 'it seems', and *credere* 'to believe', account for nearly half of all dubitative tokens and together represent 59% of all the subjunctive morphology attested in this class, reinforcing the influence of the lexical identity of the governor rather than a genuinely semantic effect, and suggesting that the subjunctive may be triggered by these specific verbs themselves rather than by their purported underlying meaning in discourse.

4.2. The structure of the matrix clause

Among the structural features of the matrix clause hypothesized to condition subjunctive selection is sentence type. If the subjunctive signals a weaker speaker commitment to the truth of the embedded proposition, as generally claimed (Carlier et al. 2012; Quer 2001; Wandruszka 1998; Loengarov 2006; Manzini 2000; Veland 1991), it should be favoured in non-declarative contexts, such as interrogatives and conditionals, which have been variously characterized as non-assertive or less assertive than their affirmative counterparts (Haspelmath 2003: 220; Givón 2018: 139). Negation, too, has long been associated with reduced assertiveness and is often thought to favour subjunctive morphology (Thompson 1998; Acquaviva 1996; Manzini 2000; Costantini 2011; Giannakidou 1995; Giannakidou & Mari 2015; Giorgi & Pianesi 1997; Loengarov 2006; Veland 1991). To test this hypothesis, each token in the dataset was coded according to the structure of the matrix clause, affirmative declarative, interrogative, or conditional. In addition, the data was analyzed according to polarity of the clause, affirmative versus negative. Following Poplack (2013, 2018), these factor groups were intended as independent approximations of the assertiveness of the matrix predication, separate from the semantic class of the governor.

Despite these predictions, the results do not lend support for the hypotheses. If anything, subjunctive is disfavoured in non-declarative contexts.

	% Subj	**N**
declarative	66%	390/592
non-declarative	58%	14/24
Total	**66%**	**404/616**

Table 4: Subjunctive rate by sentence type: declarative vs. non-declarative matrix clauses.

Now turning to the testing of affirmative versus negative contexts. Likewise, results do not lend support for the hypothesis that less-assertive contexts favour the subjunctive; if anything, the opposite holds, with affirmative contexts favouring it (72%) and negative ones highly disfavouring it (49%).

	% Subj	**N**
affirmative	72%	320/443
negative	49%	84/173
Total	**66%**	**404/616**

Table 5: Subjunctive rate by polarity of the matrix clause.

Not all governors appeared in both affirmative and negative contexts. If we isolate those that did, we find 20 such governors, listed in the table below with their rates of subjunctive selection.

Governor	**Translation**	**Neg. %Subj**	**Aff. %Subj**	**p-value**	**Sign.**
credere	to believe	73% (11/15)	64% (14/22)	0.7235	n.s.
dire	to say	100% (2/2)	13% (3/24)	-	-
pensare	to think	67% (6/9)	73% (58/79)	0.6999	n.s.
sembrare	to seem	83% (5/6)	74% (43/58)	1.0000	n.s.
volere	to want	75% (3/4)	87% (13/15)	-	-
parere	to appear	100% (1/1)	56% (14/25)	-	-
è possibile	it is possible	100% (1/1)	0% (0/2)	-	-
essere sicuro	to be sure	100% (1/1)	40% (2/5)	-	-
avere il dubbio	to have doubt	100% (1/1)	50% (1/2)	-	-
Sub-total		**83% (35/42)**	**64% (148/232)**		

Governor	*Translation*	*Neg. %Subj*	*Aff. %Subj*	*p-value*	*Sign.*
Only Neg Gov		***%Subj***			
non è	it is not	33% (37/112)			
non sapere	not to know	33% (2/6)			
non è detto	it is not certain	100% (3/3)			
non è vero	it is not true	50% (1/2)			
non dispiacersi	not to mind	100% (1/1)			
non è giusto	it is not fair	100% (1/1)			
non avere senso	not to make sense	100 % (1/1)			
non è da dire	it is not to be said	100% (1/1)			
non è umano	it is not humane	100% (1/1)			
non preoccuparsi	not to worry	100% (1/1)			
Sub-total		**61% (80/131)**			

Table 6: Subjunctive rate by polarity (affirmative vs. negative) for frequent governors. Fisher's exact test conducted only for governors with N ≥ 5 in both polarity contexts.

Eleven of these governors occurred only in negative form, all with very low token counts or as singletons. Only *non sapere* 'not to know' (12a below) appears with slightly more tokens, though still few, and disfavours the subjunctive, which is consistent with expectations that this verb is not typically associated with subjunctive use (Wandruszka 1998: 442; Schmitt Jensen 1970: 234). A substantial portion of the negative data, however, comes from a single governor, *non è* 'it is not' (12b), which is not only frequent but also strongly disfavours the subjunctive (33%).

(12) a. SAPERE 'TO KNOW'

Io non l'ho studiato perché <u>non sapevo</u> ***era*** *[3SG.IMPRF.IND] da fare.* (C.010.23)

'I didn't study it because I didn't know it had to be done.'

b. NON È 'IT IS NOT'

Ah <u>non è</u> che lo ***devo***[3SG.PRES.IND] *pagare io.* (L.120.81)

'Ah, it's not that I have to pay for it.'

As for the remaining negative contexts that appear to favour the subjunctive, the apparent effect is inflated by the presence of only one to three occurrences and is not interpretable in any meaningful sense.

If we examine the sub-group that showed variability in both affirmative and negative

governors, while subjunctive selection appears to be higher in negative clauses (83%) than in affirmative ones (64%) for these items, low token counts prevent firm conclusions. For those with ≥5 tokens in both contexts, we conducted Fisher's exact tests[7] to assess whether the observed differences were statistically significant. None reached significance at the $p < 0.01$ level (n.s.), suggesting that the presence of negation alone does not independently determine mood selection.

Overall, neither interrogative and conditional sentence types nor polarity support a semantically-driven account of mood selection. Results either run counter to expectations or implicate the same governors, *credere* 'to believe (13a), *pensare* 'to think' (13b), *sembrare* 'to seem' (13c), *volere* 'to want' (13d), suggesting that subjunctive selection is more closely tied to lexical identity than to the semantics of non-assertiveness.

(13) a. CREDERE 'TO BELIEVE'

Io credo che ***sia*** [3SG.PRES.SUBJ] *stata la meglio ditta di tutta l'Italia.* (C.322.151)

'I believe that it has been the best firm of all Italy.'

b. PENSARE 'TO THINK'

*Eh invece penso ora tu c'****abbia*** [3SG.PRES.SUBJ] *parecchia esperienza.* (C.5.110)

'Eh but I think now you have lots of experience.'

c. SEMBRARE 'TO SEEM'

*Ma veramente, sembra c'****abbia*** [3SG.PRES.SUBJ] *quattordici, quindic'anni.* (C.072.227)

'But honestly, it seems he is fourteen, fifteen years old.'

d. VOLERE 'TO WANT'

Senti, vuoi che ***compri*** [3SG.PRES.SUBJ] *i popcorn e la Coca-Cola?* (L.412.96)

'Listen, do you want me to buy popcorn and coke?'

Notably, earlier quantitative studies (e.g., Schneider 1999; Veland 1991), motivated by the assertion hypothesis, excluded negative and interrogative clauses on the grounds that these contexts categorically trigger the subjunctive. The present findings refute this assumption and highlight that such exclusions not only compromise the principle of accountability, but also preclude the possibility of empirically and systematically testing the very assumptions they seek to uphold.

[7] Fisher's exact test is preferable when working with small sample sizes or when expected frequencies in contingency table cells are very low (Gorman & Johnson 2013: 2019).

4.3. Other indicators of non-factual modality

A final test of semantics follows the principle of redundancy (Poplack et al. 2018), which posits that "any meaning expressed by the subjunctive would be echoed elsewhere in the discourse" (Poplack et al. 2018: 233). In the case of the subjunctive, often described as marking non-factual or uncertain propositions, one would expect its use to be reinforced by other contextual indicators of non-factuality, whether lexical, modal, or structural. These include adverbs such as *forse* 'maybe', *probabilmente* 'probably', valuative expressions like *magari* 'perhaps' or *speriamo che* 'hopefully', epistemic modals such as *può darsi* 'it might be', or metadiscursive constructions like *se non mi sbaglio* 'if I'm not wrong'. Other indicators comprise the use of matrix moods such as the conditional, future, imperative, or subjunctive, forms argued to mark speaker detachment from the assertion or to promote so-called "modal attraction" (Gatta 2002; Hooper 1975; Klein 1975; Loengarov 2006; Manzini 2000; Santulli 2009; Veland 1991; Wandruszka 1991).

According to these accounts, such elements are sometimes described as 'subjunctivizing' factors or triggers of the subjunctive, and have occasionally been treated as categorical contexts for its selection. From a variationist standpoint, however, such assumptions are tested empirically. The current study therefore codes for the presence of explicit indicators of non-factual modality and assesses whether they in fact favour subjunctive selection.

Following Poplack (2013; 2018), these indicators were operationalized as a factor group distinct from sentence type and from the lexical identity of the embedded predicate. In keeping with the principle of orthogonality (Guy 1988: 126), a context already captured in another factor group, such as conditional sentences, was excluded here to avoid redundancy and confounding effects. That is, although conditional clauses may structurally promote a non-factual interpretation and should theoretically favour the subjunctive (Manzini 2000: 243), they were already analyzed in the "sentence type" factor group (Section 4.2), and are therefore not repeated here. This approach isolates the effect of non-factual indicators not otherwise captured, allowing for a more reliable test of the semantic conditioning hypothesis.

(14) a. MOOD OF THE MATRIX CLAUSE

Ecco, a me piacerebbe che 'un ci ***fosse***$_{[3SG.IMPRF.SUBJ]}$ *questa-questo comportamento schizofrenico.* (C.430.241)

'So, I would like it not to be this-this schizophrenic behaviour.'

b. AUXILIARY USED MODALLY

Uno potrebbe pensare che architettura e struttura sono$_{[1SG.PRES.IND]}$ *la stessa cosa.* (C.451.25)

'One could think that architecture and structure are the same concept.'

c. LEXICAL

Magari ci poteva stare che io andassi$_{[1SG.IMPRF.SUBJ]}$ *a cambiarli.* (C.224.61)

'Maybe it made sense that I could go to exchange them.'

	% Subj	**N**
Presence	72%	55/76
Absence	65%	349/540
Total	**66%**	**404/616**

Table 7: Subjunctive rate according to the presence of other indicators of non-factual modality.

A first observation is that indicators of non-factual modality are relatively rare in running discourse, with only 12% of tokens enabling us to test whether their presence affects mood selection. According to our hypothesis, one would expect that any meaning conveyed by the subjunctive would be echoed by other contextual cues contributing to a non-factual reading. Indeed, results show a higher rate of subjunctive in the presence of such indicators (72%) compared to their absence (65%). However, this difference is not statistically significant at $p < 0.01$ (χ^2, $N = 616$) $= 1.44$, $p = 0.23$), suggesting that these elements do not reliably condition the selection of the subjunctive. Moreover, if we break down the category according to the specific type of indicator coded for, results show that not all indicators behave uniformly: modal auxiliaries categorically trigger the subjunctive; among the two categories that account for most of the data of indicators of non-factual modality, tense and mood of the matrix clause favour the subjunctive (77%), while the presence of lexical indicators has no effect (66%).

Type of Indicator	**% Subj**	**n Subj**
Auxiliary used modally	100%	7/7
Combination of factors	100%	2/2
Mood of the matrix clause	77%	17/22
Lexical	66%	31/47
Absence	64%	347/538
Total	**66%**	**404/616**

Table 8: Subjunctive rate according to various indicators of non-factual modality.

Summarizing, factors designed to capture a non-factual or doubtful reading through explicit cues in discourse do not consistently influence subjunctive selection, contra the hypothesis of semantic contribution. If anything, the only context where the effect is consistent is with modal auxiliaries, though low token counts preclude firm conclusions. These findings corroborate those for the other semantics-based factor groups: if semantics were genuinely explanatory of subjunctive variability, we should not observe such inconsistency both across and within factor groups. Of the three groups tested, none accounted for a semantic contribution. Instead, results suggest that a handful of governors account for 47% of the data and 56% of all subjunctive morphology: *pensare* 'to think' (73%), *sembrare* 'it seems' (75%), *credere* 'to believe' (79%), *bisognare* 'it is necessary' (93), *volere* 'to want' (84%), and *sperare* 'to hope' (93). Each favours the subjunctive, and moreover, these are the only verbs that show a higher proportion of subjunctive morphology than their share of the data, suggesting that subjunctive selection is more closely tied to lexical identity than to the semantic context (see Appendix).

This trend was observed also among dubitative governors (*pensare* 'to think', *sembrare* 'it seems', and *credere* 'to believe') and volitive predicates (*volere* 'to want', *bisognare* 'it is necessary', and *sperare* 'to hope') when we tested the semantic category of the main predicate. While the semantic classification of these predicates on the surface predicts the expected outcomes, with volitive favouring more than other classes the selection of the subjunctive, albeit not categorically so in our dataset, the high variability in terms of rates of subjunctive usage across governors and the dominant presence of just a few verbs within each semantic class suggest that the effect is less a function of semantic class per se and more attributable to these specific verbs.

5. Diachronic perspective on subjunctive selection

5.1. A glimpse into subjunctive selection in Italian: from the 16th century to contemporary speech

Previous variationist research examined the use of the Italian subjunctive across five centuries (Digesto 2019), particularly scrutinizing completive clauses to assess whether its use was productive as well as semantically motivated in earlier stages of the language. Digesto (2019) utilized written sources to examine historical stages of Italian, alongside recordings of contemporary spontaneous speech[8]. The historical benchmarks were established based on a corpus of theatrical scripts, predominantly comedies. These texts, while written, serve as proxies for spoken Italian across time, embodying features indicative of oral communication (Stefanelli 2006: 57), alongside documented vernacular features such as mood alternation under identical matrix verbs (15), exemplified by *credere* 'to believe'; the use of the adverb of negation *mica*, "characteristic of colloquial, informal, registers" (16) (Maiden & Robustelli 2007: 405); the indirect object pronoun *gli*$_{3SG.M}$ (masculine) instead of *le*$_{3SG.F}$ (feminine) with 3rd person singular feminine indirect objects (17).

(15) *Possono loro credere sul serio che la mia figliuola* ***è***$_{[3SD.PRES.IND]}$ *morta? Che io* ***sia***$_{[3SD.PRES.SUBJ]}$ *pazza? Che questa che ha con sé* ***è***$_{[3SD.PRES.IND]}$ *una seconda moglie?* (COHI.20C.098.754)
'Can they really believe that my daughter is dead? That I am crazy? That the one who is with him is his second wife?'

(16) *Non ho mica voluto io che s' arrivasse a una cosa così grave.* (COHI.20C.072.2253)
'I didn't really want to come to such a dangerous point.'

(17) *Gli*$_{[3SG.M]}$ *ho dato speranza di condurla, ancor oggi, alle voglie sue.* (COHI.18C.033.280)
'I gave her hope that I will persuade her again today.'

[8] Contemporary speech data comes from LIP (De Mauro et al. 1993) and C-ORAL-ROM (Cresti & Moneglia 2005), while the theatrical plays selected as a benchmark for the historical data were compiled as part of the COHI, *Corpus of Historical Italian* (Digesto 2019). We refer the reader to Digesto (2019; 2021) for detailed description of the corpora.

The findings (2019) indicate a relative stability in the patterns of usage of the subjunctive in Italian over time. Despite its seeming productivity, as it remains the dominant variant in completive clauses both in contemporary (66%) and written historical Italian (79% in the 16th century; 81% in the 18th century), its frequent usage is attributed to a limited number of lexical governors.

	COHI			SPEECH
	16th Century	**18th Century**	**20th Century**	**20th/21st Century**
%Subj	79%	81%	74%	66%
n Subj	299/379	408/505	207/279	404/616
n words	19,621	37,264	40,565	387,825
n governors	81	88	60	71

Table 9: Overall rate of subjunctive selection (%Subj), number of subjunctives (*n* Subj), number of governors that occurred with a subjunctive (*n* governors), and total number of words (*n* words) for each time period.

These trends hold despite the variances in the nature of the data, timespan, conversational topics, and the social backgrounds of both speakers in spontaneous interactions and characters in the comedies. It has been noted that a small subset of verbs such as *credere* 'to believe', *pensare* 'to think', as well as *sembrare* 'it seems' in contemporary speech accounts for the majority of the variation observed in discourse; that the subjunctive is likely to occur with suppletive forms of *essere* 'to be'; and that dedicated tests of semantic contribution did not substantiate the theory of a semantically-driven subjunctive, neither historically nor in present-day use. This evidence supports the premise that a lexicalization pattern has been operative since at least the 16th century.

In examining specific semantic categories (2019: 188), such as *necessity*, *volition*, and *emotion*, which are theoretically categorical triggers of subjunctive mood due to their strong modal determination, results revealed variability with indicative mood yet remaining strong predictors for the selection of the subjunctive over time. Although the effect of necessity verbs appeared as an epiphenomenon of a lexical effect due to *bisognare* 'it is necessary', *emotive* and *volitive* verbs demonstrated a more consistent effect. Despite this consistent selection, often at high rates, occasionally categorically, suggesting semantic motivation at least for these contexts, the rarity of

these governor types within the variable context and the disproportionate frequency of certain governors within the same semantic category, especially among volitive verbs, complicate the picture. Notably, within the *volitive* semantic class, the verb *volere* 'to want' singularly represented a significant proportion of usage, constituting 40% in the 16th century, 57% in the 18th century, and 73% in the 20th century, casting doubt on a purely semantic interpretation and suggesting a more lexically driven selection of the subjunctive.

5.2. Tracing complementation patterns from Vulgar Latin to Italian

Considering the aforementioned distinctions between Latin and its Romance descendants, particularly Italian, regarding 1) the preference for finite versus non-finite constructions according to the main clause predicate and 2) the shift from left- to right-branching structures, and considering the way that the variable context for subjunctive has been defined, syntactic complementation differences are likely to influence our findings. Initially, we identified subjunctive completive clauses within the Latin corpus using the methodology outlined above. To determine the continuity of subjunctive governors across languages and through time, we examined each governor identified in VL within the Italian dataset, noting the absence or presence of these governors in the Italian context, and vice versa. This approach helped to ascertain the extent of governor transmission from Latin to Italian, their ongoing association with the subjunctive mood, and the nature of transmitted conditioning.

However, the potential impact of Latin's distinct syntactic strategies on our analysis remains a concern. Specifically, we may encounter two scenarios: a) certain governors might not be identified within the variable context if they did not select a finite complement clause, hence no subjunctive usage; b) some governors in our dataset may exhibit a preference for non-finite complementation, leading to infrequent subjunctive selection (Figure 2). To better understand if a tendency for non-finite complementation in VL influenced the selection of subjunctive or indicative mood in completive clauses, we expanded the variable context for VL. This included examining every governor mentioned in Petronius's *Cena* and those found in Italian but not in VL, independently of their frequency of selecting non-finite complements. Such an approach nonetheless enables more reliable results on the use of subjunctive in VL's completive clauses and a deeper comprehension of governor persistence across languages and through time.

5.3. The overall trajectory of the subjunctive

In the VL data, 20 distinct governors were identified with an overall near-categorical rate of subjunctive selection: 97% (71/73).

Governor	Translation	%Subj	N
rogo	to ask	100%	22/22
curo	to ensure	89%	8/9
suadeo	to recommend	100%	6/6
timeo	to fear	100%	5/5
oportet	it is necessary	100%	5/5
caveo	to beware	100%	3/3
licet	it is permitted	100%	3/3
persuadeo	to persuade	100%	3/3
nolo	not to want	100%	2/2
non est	it is not	100%	2/2
spero	to hope	50%	1/2
dubito	to doubt	100%	2/2
video	to understand	100%	2/2
miror	to be astonished	100%	1/1
experior	to find out	100%	1/1
exspecto	to expect	100%	1/1
volo	to want	100%	1/1
veto	to forbid	100%	1/1
indignatus est	to be resented	100%	1/1
efficio	to make	100%	1/1
Total		**97%**	**71/73**

Table 10: Subjunctive selection with verbal governors in Vulgar Latin ('Cena Trimalchionis' from the *Satyricon* by Petronius). Governors are sorted by total number of tokens.

The indicative only appears in two tokens, one with the governor *spero* 'I hope' (18) and one with *curo* 'I care' in its negative form (19).

(18) ***Spero*** *tamen iam veterem pudorem sibi imponit*$_{[2SG.PRES.IND]}$. (VLCo.XLVII.492).
'I hope at least that my stomach imposes some decencies on itself.'

(19) *Nam quod strabonus est*$_{[3SG.PRES.IND]}$ ***non curo.*** (VLCo.XLVII.492)
'For instance, that he be crossed-eyed, I don't care.'

The initial findings highlight a notable lack of variability in the use of governors in VL. Moreover, the low number of governors contrasts with what has been found diachronically in Italian: by the 16th century, the earliest period examined in variationist research, Italian already exhibited 81 governors, highlighting a dynamic expansion. Despite this growth in Italian, the role of governors in VL regarding subjunctive selection mirrors earlier observations, emphasizing their influence over the quantity of available data. In other words, more data does not entail more subjunctives. The *Cena Trimalchionis* contributes a corpus of 12,576 words, comparable in size to the 16th-century Italian corpus (19,621 words; 81 governors), yet VL demonstrates a much smaller pool of governors (N = 20) and a limited number of occurrences for nearly all (totaling only 73). This suggests that complement clauses governed by a verb were not a highly productive context for the subjunctive in VL, unlike the trend seen in Italian.

Exploring the continuity of governors over time, the persistence of governors was examined, which enables us to establish stability and change with respect to the governors when these are coupled with their relative rates of occurrence and frequency within the variable context. Results reveal initial insights into lexical transmission. The near-total overlap of governors between VL and Italian (19 out of 20) preliminarily indicates lexical continuity. However, this figure pales in comparison to the 239 lexical types identified in Italian across the centuries (Digesto 2019: 214). Investigating these Italian governors within the Latin dataset, only one, *credo* ('I believe'), was found not to select the subjunctive in VL, indicating a potential reason for the limited number of governors in VL: they might not have constituted subjunctive-selecting contexts historically. Except for *indignatus est* 'to be resented', all other 19 governors identified in the VL data were transmitted to Italian, with a third of them (N = 6/20) persisting through all five examined periods. These include *exspecto* 'I expect', *dubito* 'I doubt', *volo* 'I want', *nolo* 'I do not want', *oportet* 'it is necessary', and *credo* 'I believe', though *credo* uniquely favoured the indicative in VL, pointing to nuanced shifts in mood selection over time.

The disparity in the number of governors and the scant occurrence of tokens in VL as compared to Italian necessitates further exploration, particularly in light of the previously discussed potential syntactic divergences in complementation. This discrepancy prompts an inquiry into whether the distinct syntax between the languages, specifically the prevalent use of infinitival clauses in VL, including those with an accusative subject (as noted by Bolkestein 1989 and Bodelot 2003), could account for the limited presence of finite complement clauses and, consequently, the sparse subjunctive data in VL. To

determine if the linguistic scenario observed with verbal governors in Latin primarily stems from profound syntactic differences, an analysis was conducted on each Latin governor's complementation type.

	Vulgar Latin		**16th Century**		**18th Century**		**20th Century**		**Actual Speech**	
	%	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**
Sharedness = 5										
nolo (≃non volere)	100	2/2	100	5/5	100	26/26	100	3/3	75	3/4
volo (≃volere)	100	1/1	100	22/22	100	47/47	100	49/49	87	13/15
oportet (≃bisognare)	100	5/5	100	1/1	100	29/29	100	9/9	92	35/38
dubito (≃dubitare)	100	1/1	100	1/1	75	3/4	67	2/3	50	1/2
exspecto (≃aspettare)	100	1/1	100	3/3	88	7/8	100	1/1	50	4/8
credo (≃credere)	0	0/6	83	39/47	98	42/43	69	22/32	79	45/57
Sharedness = 3										
rogo (≃pregare)	100	16/16	100	8/8	100	3/3				
video (≃guardare)	100	1/1	100	3/3	100	3/3				
suadeo (≃consigliare)	100	6/6	100	1/1	100	1/1				
Sharedness = 2										
exspecto (≃attendere)	100	1/1	100	1/1						
Sharedness = 3										
non est (≃non è)	100	2/2	100	1/1					33	37/112
esperior (≃scoprire)	100	1/1			100	1/1				
caveo (≃badare)	100	3/3			80	4/5	0	0/1		
rogo (≃chiedere)	100	16/16			100	1/4			66	2/3
Sharedness = 4										
miror (≃meravigliarsi)	100	1/1	100	1/1	100	4/4				
timeo (≃consigliare)	100	5/5	75	3/4	100	2/2				
curo (≃assicurare)	100	7/7			33	2/2	0	0/6		
timeo (≃avere paura)	100	5/5			100	7/7	100	1/1	75	3/4
licet (≃permettere)	100	3/3			100	6/6	100	2/2		
spero (≃sperare)	50	1/2			38	3/8	50	1/2	93	14/15
efficio (≃fare)	100	1/1	100	16/16	100	7/7			71	5/7
Sharedness = 2										
exspecto (≃attendere)	100	1/1	100	1/1						
veto (≃vietare)	100	1/1			100	1/1				

Table 11: Persistence of governors across time periods (from 2 to 5) and rate of subjunctive selection. Governors are paired according to roughly the same meaning. The columns regarding 16th, 18th and 20th century data report the results of the corpus of Italian comedies over time.

As indicated earlier (Figure 3), there is a suggestion that Latin favoured non-finite complements more so than its Romance successors (as per Magni 2009, among others). It is reasonable to hypothesise that ongoing syntactic shift from OV to VO structures in Latin, alongside a continued preference for infinitival constructions, results would manifest markedly different strategic choices compared to Italian, which is predominantly an SVO language. Should a preference for non-finite complementation obstruct the emergence of the subjunctive in VL, this would elucidate the observed low productivity in the ancestral language and provide concrete evidence of a fundamental syntactic disparity between VL and Italian.

TYPE OF COMPLEMENT:		FINITE						NON-FINITE		TOTAL	
		SUBJS		IND		AMBIGUOUS					
GOVERNOR		**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**
rogo	to ask	100%	22	0%	0	0%	0	0%	0	0	22
curo	to ensure	89%	8	11%	1	0%	0	0%	0	0	9
suadeo	to recommend	100%	6	0%	0	0%	0	0%	0	0	6
timeo	to fear	100%	5	0%	0	0%	0	0%	4	0	5
oportet	it is necessary	56%	5	0%	0	0%	0	44%	0	15	9
caveo	to beware	75%	3	0%	0	25%	1	0%	0	0	4
licet	it is permitted	43%	3	0%	0	57%	4	0%	0	0	7
persuadeo	to persuade	43%	3	0%	0	57%	4	0%	4	0	7
nolo	not to want	33%	2	33%	1	0%	0	67%	0	15	6
non est	it is not	100%	2	0%	0	0%	0	0%	0	0	2
dubito	to doubt	100%	2	0%	0	0%	0	0%	6	0	2
video	to understand	25%	2	0%	0	0%	0	75%	1	23	8
spero	to hope	33%	1	0%	0	0%	0	33%	0	4	3
miror	to be astonished	100%	1	0%	0	0%	0	0%	0	0	1
efficio	to make	100%	1	0%	0	0%	0	0%	0	0	1
indignatus est	to be resented	100%	1	0%	0	0%	0	0%	0	0	1
experior	to find out	100%	1	0%	0	0%	0	0%	0	0%	1
exspecto	to expect	100%	1	0%	0	0%	0	0%	11	0	1
volo	to want	8%	1	0%	0	0%	0	92%	0	42	12
veto	to forbid	100%	1	0%	0	0%	0	0%	0	0	1
credo	to believe	0%	0	86%	6	14%	1	0%	0	0	7
Total		**62%**	**71**	**7%**	**8**	**9%**	**10**	**23%**	**26**		**115**

Table 12: Number of tokens (N) and proportion of finite vs. non-finite complementation (%) for each verbal governor in Vulgar Latin data. Governors are sorted by total number of tokens with subjunctive morphology.

Only five out of twenty-one governors opted for infinitival clauses over finite subordinate ones, with *oportet* 'it is necessary', *nolo* 'I do not want', *volo* 'I want', *spero* 'I hope', and *video* 'to understand' showing a preference for non-finite constructions. A significant portion of these non-finite contexts is dominated by *volo* 'I want' (42%) and further increased by *nolo* 'I do not want', together accounting for 58% of cases eschewing subjunctive complements. Apart from these, the larger part of governors (16/21) exclusively chose finite complements. The difficulty in differentiating subjunctive from indicative moods due to morphological similarities in Latin might have influenced our results, suggesting an underestimation of subjunctive productivity.

Given the high degree of morphological homophony between the subjunctive and the indicative in Latin, our results had a relatively greater likelihood to be affected by such cases, preventing us from distinguishing the embedded mood and generating the apparent non-productivity observed here. However, ambiguous morphology was not a significant factor in limiting the diversity of subjunctive governors, as only three governors out of the 20 subjunctive-selecting contexts showed such ambiguity: *caveo* 'to beware', *licet* 'it is permitted', and *persuadeo* 'I persuade'. Contrary to expectations of widespread infinitival usage, our findings suggest that the choice between finite and non-finite complements does not explain the differences between VL and Italian, particularly regarding the expansion and variability of subjunctive selecting governors in Italian. The predominance of finite complements (89 tokens) over non-finite ones (26 tokens) among our identified governors suggests a preference for *finite* complementation in VL.

6. The grammaticalization path of the subjunctive

6.1. Subjunctive use from Latin to Italian

Although VL presented a limited dataset, notable patterns emerged concerning the use of the subjunctive in verb-governed completive clauses across VL and Italian. First of all, there is superficial evidence of linguistic change; specifically, the overall rate of subjunctive selection decreases from a quasi-categorical 97% in VL to 66% in contemporary Italian speech, albeit reaching 81% in the 18th century (81%).

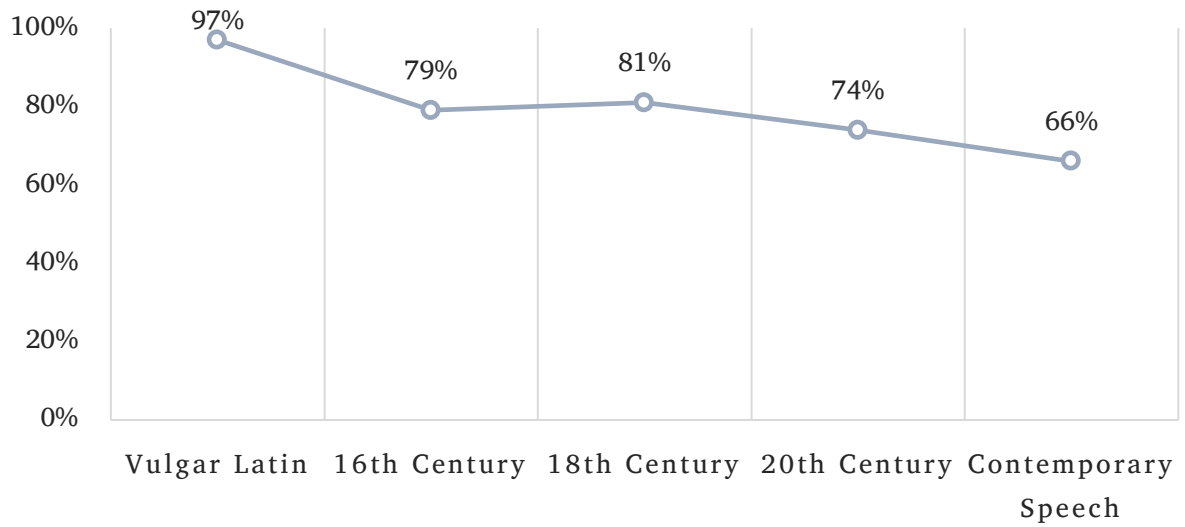


Figure 4: The trajectory of subjunctive selection across time from Vulgar Latin to 20th/21st century Italian.

The comparison between VL and Italian highlights a significant discrepancy in the number of governors, showcasing VL's relative unproductivity in this aspect, especially when juxtaposed with the more expansive governor pool in contemporary Italian. Despite analyzing a VL corpus of similar size to that of 16th-century Italian, VL exhibits a much smaller set of governors, totaling only 20 verbs, indicating a stark contrast across the languages. This observation raises questions about the dynamics of linguistic change, particularly why there is such a variance in governor numbers between VL and Italian. A deeper analysis into the application of subjunctive morphology across various completive clauses offers additional insight. The systematic coding and examination of subjunctive use within both VL and Italian contexts reveal a distinct split: in Italian, subjunctive morphology increasingly appears with verbal governors, whereas in VL, it is predominantly associated with non-verbal governors, especially conjunctions like *ut* 'so that' in VL (and its Italian counterpart *in modo che* 'so that').

(20) *Ego gloriosus volo efferri* ***ut*** *totus mihi populus bene imprecetur*$_{SBJV}$.
(VLCo.LXXVIII.1251).
'I want to be carried out in splendour, so that the whole crowd calls down blessings on me.'

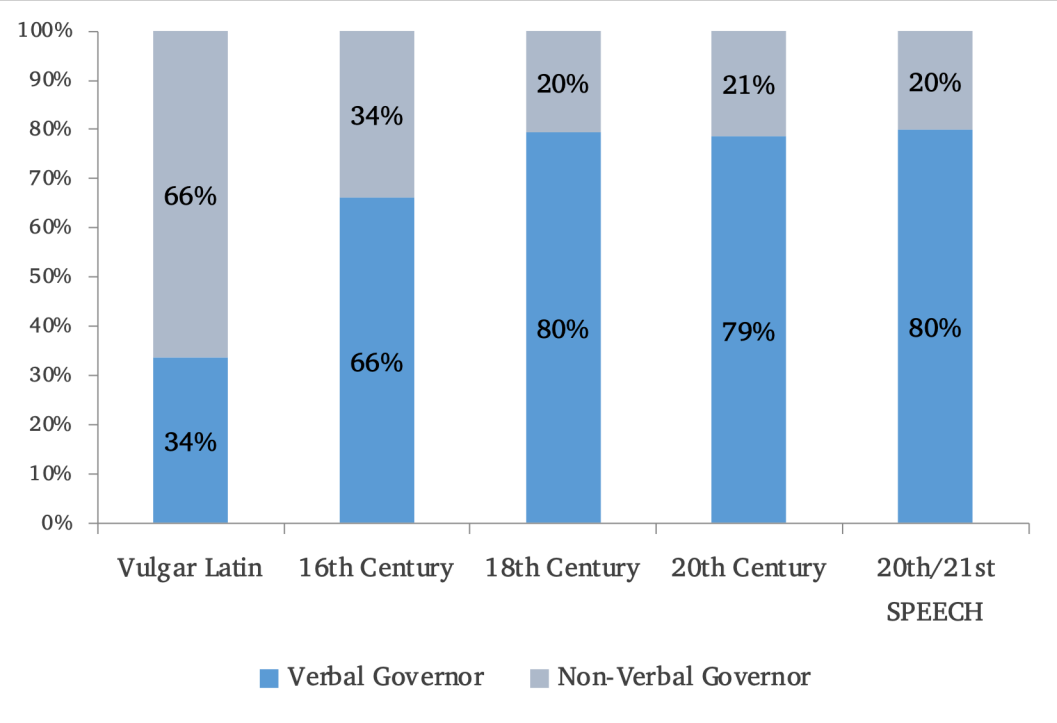
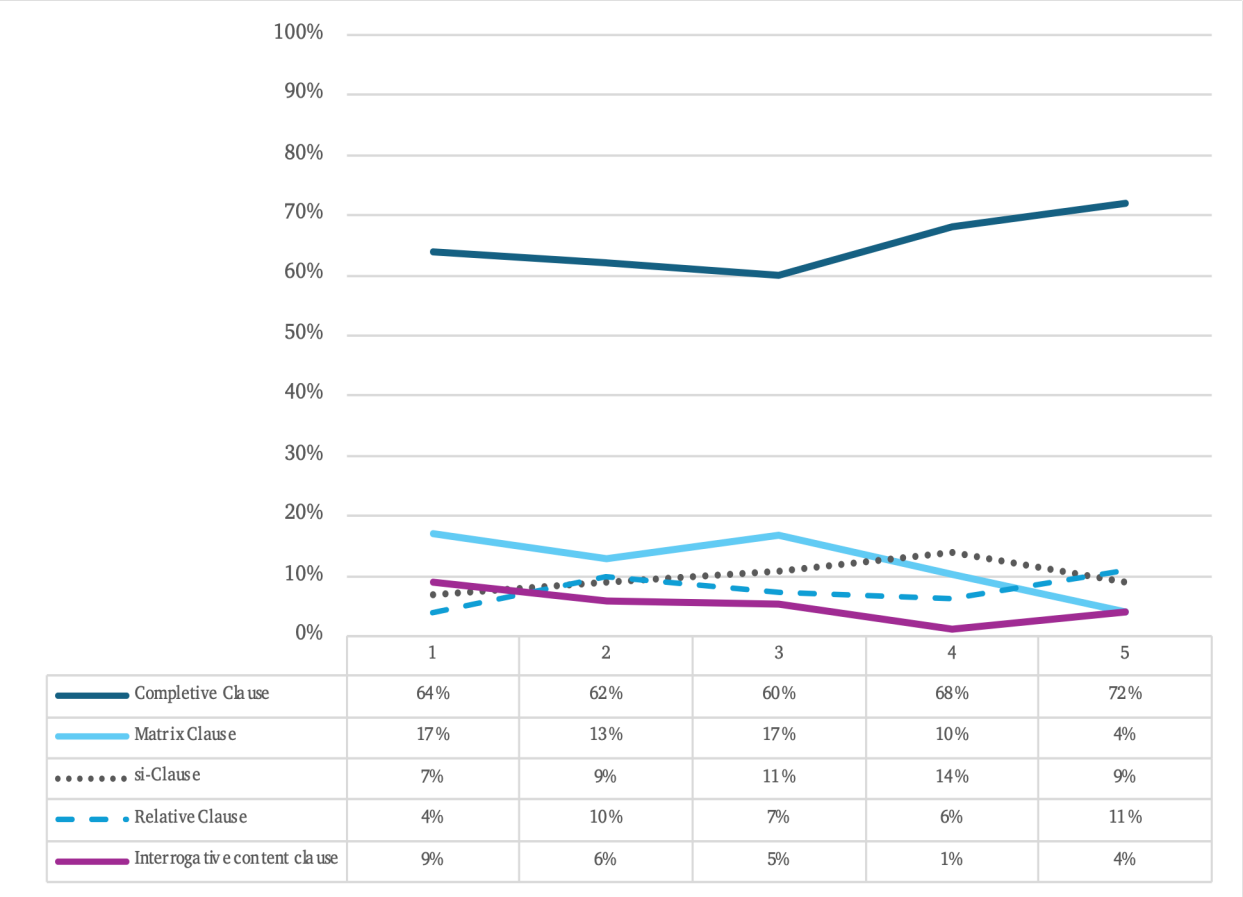


Figure 5: Distribution of subjunctive morphology in the context of completive clauses across time (from Vulgar Latin to contemporary Italian speech) when governed by either a verbal governor (object of study of the current research) or by a non-verbal governor.

This divergence underscores a fundamental shift in the usage of the subjunctive, suggesting that while in Latin, the subjunctive's association with conjunctions might have been reinforced by the syntactic transition from OV to VO, facilitating primarily finite complementation, Italian displays the opposite trend. This shift becomes evident in the 16th century, with a significant percentage (66%) of subjunctives governed by verbs, a figure that solidifies in the 18th century at 80%. Analyzing the distribution of subjunctive morphology across different contexts over time illuminates the evolving patterns of its usage.

Consistent with earlier research (Harris 1978; Noonan 2007), subjunctive morphology in VL *primarily* marked subordination, particularly within completive clauses, indicating its established role as a subordination marker. The distribution analysis reveals its limited use in independent clauses, showing a decline over time from 17% in VL to 4% in contemporary Italian. An overall decreasing trend is observed in the context of interrogative content clauses, from 9% in VL to 4% in contemporary Italian. Subjunctive use within *if*-clauses has remained relatively consistent, while relative clauses see an upward trajectory. The distribution pattern indicates a stronger association of subjunctive morphology with completive clauses in both Latin and Italian, suggesting this syntactic context as the primary domain for

subjunctive usage, a trend that has also strengthened from 64% in VL to 72% in modern Italian.



	1	2	3	4	5
Completive Clause	64%	62%	60%	68%	72%
Matrix Clause	17%	13%	17%	10%	4%
si-Clause	7%	9%	11%	14%	9%
Relative Clause	4%	10%	7%	6%	11%
Interrogative content clause	9%	6%	5%	1%	4%

Figure 6: Distribution of subjunctive morphology across syntactic environments, from Vulgar Latin to contemporary Italian speech.

Figure 5 clearly shows that the subjunctive had *already* spread out in VL to other subordinating contexts, such as relative clauses, indirect question clauses, etc. (as pointed out by Harris 1974, 1978; Magni 2009; Haudry 1973), which supports the notion that a *generalization* had already occurred by that time. Its extended use in subordinate syntactic position can be taken as evidence of the subjunctive being interpreted as a *subordinate marker* and a shift from a stronger to a weaker hypotactic link between the clauses (Magni 2009). This extended subordinating use in Latin, and the subsequent transmission of its subordinating function to Italian, along with an expanded governor pool, highlights the continuity and evolution of the subjunctive's role across languages.

6.2. Semantic factors over time

The only result landing support to a supposed semantic conditioning in Italian was related to quasi-categorical selection of subjunctive mood with verbs expressing volition, which remains stable over time. When we examine the semantic class of the VL subjunctive governors, the semantic class with the richest lexical inventory is indeed that of verbs marking volition (85% of the data).

Semantic Class	Governor	Translation	%	N
Volitive	*rogo*	to ask	100%	22/22
	curo	to ensure	89%	8/9
	suadeo	to recommend	100%	6/6
	timeo	to fear	100%	5/5
	oportet	it is necessary	100%	5/5
	caveo	to beware	100%	3/3
	licet	it is permitted	100%	3/3
	persuadeo	to persuade	100%	3/3
	nolo	not to want	100%	2/2
	spero	to hope	50%	1/2
	exspecto	to expect	100%	1/1
	volo	to want	100%	1/1
	veto	to forbid	100%	1/1
	efficio	to make	100%	1/1
	Sub-Total		**97%**	**62/64**
Dubitative	*video*	to understand	100%	2/2
	non est	it is not	100%	2/2
	dubito	to doubt	100%	2/2
	experior	to find out	100%	1/1
	Sub-Total		**100%**	**7/7**
Factive	*miror*	to be astonished	100%	1/1
	indignatus est	to be resented	100%	1/1
	Sub-Total		**100%**	**2/2**
	TOTAL		**97%**	**71/73**

Table 13: Governors and their rates of subjunctive selection according to semantic class in VL data.

Subjunctive morphology in Latin was predominantly associated with volitive contexts, confirming the notion that these verbs primarily govern the use of the

embedded subjunctive, aligning with its original paratactic function. This observation supports the idea that the syntactic patterns of harmonic hypotactic contraction observed in Latin have been preserved and transmitted to Italian, demonstrating a continuity and inheritance within the linguistic history of the languages (Labov 1989). If we assume that the contemporary Italian subjunctive kept the semantic motivation in this context, we nonetheless need to acknowledge the fact that it is *also* found in other non-harmonic contexts such as those with other types of governors, e.g., *experior* 'to find out', albeit to a lesser extent, and is also used in other embedded complements, such as interrogative-content clauses.

Thus, the subjunctive's role in Latin and its evolution in Italian can be viewed through both a semantic lens, retaining meanings from paratactic origins, and a syntactic and lexical perspective, highlighting a continuing lexical pattern but also adaptability to new lexical governors over time.

Despite the predominance of volitive governors as triggers, the subjunctive also fulfills a broader subordinating function, a dual characteristic that has been carried over into Italian. Although it had already generalized as a marker of subordination and spread out to a variety of embedded contexts, the Latin subjunctive was mainly used with volitive verbs within *that*-clauses (Figure 7).

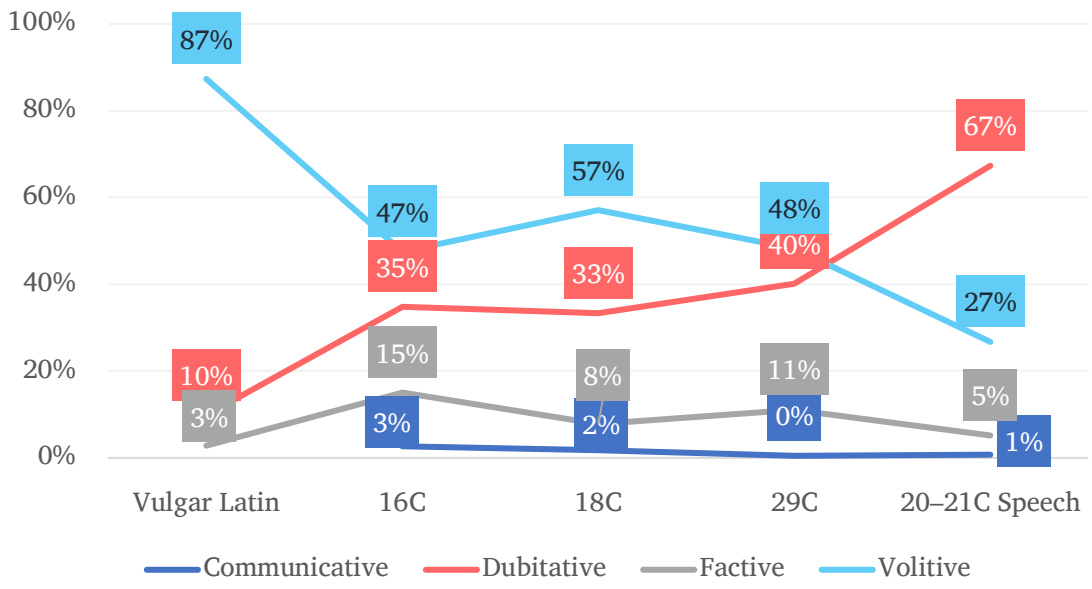


Figure 7: Distribution of subjunctive morphology across semantic classes in Vulgar Latin, historical, and contemporary Italian data.

Regarding the distinction in triggering contexts between VL and Italian, communicative and opinion verbs in VL typically lead to non-finite complement clauses, contrasting with the expanded governor pool in Italian. This would explain the important discrepancy observed with respect to the Italian situation, i.e. increased number of governors[9]. On the other hand, the strong association of the subjunctive with volitive verbs can be considered a pattern transferred from Latin into Italian, and one which is remarkably stable across centuries. We also observe an innovative trend: subjunctive morphology is gaining ground in other contexts that are non-harmonic with the original Latin's (e.g., *dubitative verbs* in the graph), providing additional support for the assumption that subjunctive use has a "relative stability so often found in Romance […] but ha[s] ceased to have any semantic import" (Harris 1974: 175).

When we examine other factor groups designed to gauge the contribution of semantics, we find that highly stable results across time with respect to the presence of other indicators of non-factual modality, which consistently favour subjunctive selection. However, as previously shown (see Digesto 2019: 197), multivariate analysis revealed that, with the exception of the 16th-century data, this factor group does not reach statistical significance, a result corroborated by the chi-square tests presented in Table 14.

Period	% Subj *presence*	% Subj *absence*	Total	χ^2 (df = 1)	p-value
16C	91% (126/138)	72% (173/241)	299/379	20.079	< 0.001 ***
18C	83% (154/186)	80% (254/319)	408/505	0.762	0.38 (n.s.)
20C	76% (51/67)	74% (156/212)	207/279	0.171	0.68 (n.s.)
20-21C Speech	72% (55/76)	65% (349/540)	404/616	1.768	0.18 (n.s.)

Table 14: Subjunctive rates according to presence of other non-factual indicators in Italian across time.

By contrast, the only finding lending support to the semantic hypothesis concerns the structure of the matrix clause: non-declarative sentences favour the subjunctive across the diachronic data and show a statistically significant effect (Digesto 2019: 197). Yet, since no such effect is observed in contemporary speech, this could suggest the erosion of semantic conditioning in present-day Italian. Still, this is an isolated result, and the broader pattern, including robust lexical effects, militates more strongly against a semantically-based analysis.

[9] It is worth noting that only a few governors account for more and more of the variation in discourse in Italian across time, namely *credere* ('to believe'), *pensare* ('to think') and *sembrare* ('to seem').

Period	% Subj *non-declarative*	% Subj *declarative*	Total	χ^2 (df = 1)	p-value
16C	96% (72/75)	75% (227/304)	227/304	16.43	5.0e-05 ***
18C	95% (138/146)	75% (270/359)	270/359	24.94	5.9e-07 ***
20C	81% (79/98)	71% (128/181)	128/181	3.25	0.071 (n.s.)
20-21C Speech	58% (14/24)	66% (390/592)	390/592	0.58	0.446 (n.s.)

Table 15: Subjunctive rates according to declarative and non-declarative structure of the matrix clause in Italian across time.

Likewise, the effect of polarity on subjunctive selection is consistent over time, with negative polarity favouring the selection of the subjunctive in diachronic data. The hierarchy is reversed in contemporary data.

Period	% Subj Aff.	% Subj Neg.	Total	χ^2 (df = 1)	p-value
16C	77% (260/339)	98% (39/40)	299/379	9.30	< 0.01 **
18C	78% (339/432)	95% (69/73)	408/505	10.36	< 0.01 **
20C	74% (183/246)	73% (24/33)	207/279	0.04	0.84 (n.s.)
20-21C Speech	72% (320/443)	77% (47/61)	367/504	0.63	0.43 (n.s.)

Table 16: Subjunctive rates according to affirmative and negative polarity across time. *For the 20–21C Speech data, the rate for negative polarity excludes tokens with the lexical governor *non è* 'it is not', which categorically disfavours the subjunctive and skews the overall distribution.

However, as shown in section 4.2 above, *non è* 'it is not' alone was responsible for lowering the overall rate of negative polarity to 49%; if we exclude the 37/112 tokens of *non è* 'it is not', the rate rises to 77% (47/61), patterning exactly like previous stages of the language. Nonetheless, recall that close inspection showed that where data was available for analysis, no significant difference was observed. In earlier periods (16C, 18C), negative contexts clearly favour the subjunctive, and the difference is statistically significant ($p < 0.01$). In the 20C dataset, however, the effect disappears entirely, and in contemporary speech (20–21C), the apparent disfavouring effect is driven almost exclusively by a single lexical item (*non è* 'it is not'), which categorically disfavours the subjunctive. Once these tokens are removed, no significant effect of polarity remains at $p < 0.01$, as shown in table 16 above. This bolsters previous observations that polarity is not an independent conditioning factor in mood selection, but rather reflects the influence of specific lexical governors.

6.3. Unveiling the path to lexicalization: insights across Romance languages

The variationist method has been employed to study subjunctive selection in completive clauses across Romance languages, including French, Italian, and Portuguese. This approach, building on the methodology developed by Poplack for studying the use of French subjunctive (1992) and leveraged in this study, has been applied to contemporary spoken Romance language corpora, facilitating a synchronic comparison (Poplack et al. 2018). Drawing from the findings of Poplack and her associates (2018), we note that subjunctive-selecting governors in the VL benchmark predominantly use the subjunctive, contrasting with the robust variability observed in all the descendant languages. Moreover, some governors select the subjunctive frequently, while others do so less often : "just a few of them make up a disproportionately large part of the entire governor pool, and the rest are very rare." (Poplack et al. 2018: 239). Although a wide range of governors can trigger the subjunctive mood, only a select few contribute *significantly* to the observed variation in discourse.

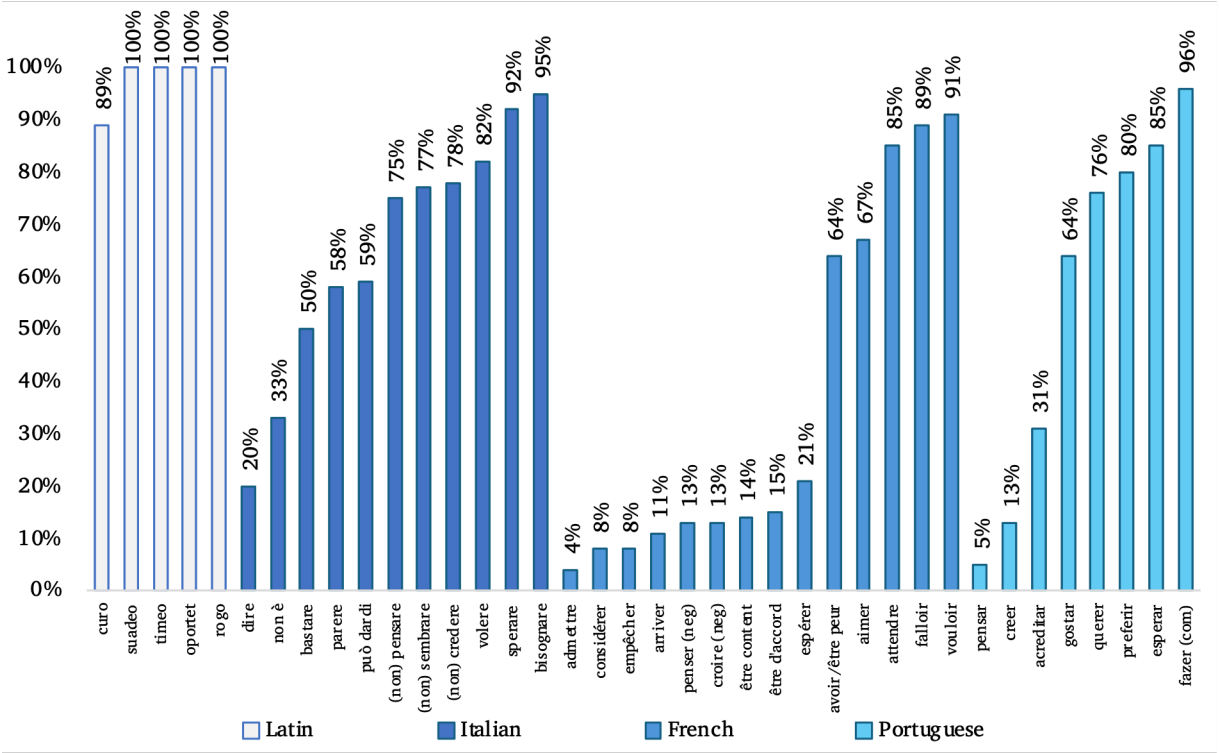


Figure 8: Subjunctive rate across frequent governors by language, adapted from Poplack et al. (2018: 239).

In a study of a 2 million-word spoken French corpus from Ottawa-Hull, Canada (Poplack 1989), *falloir* 'it is necessary' and *vouloir* 'to want' were found to represent 89% of all subjunctive morphology in complement clauses. In Brazilian Portuguese, *querer* 'to want' accounts for approximately a quarter of the governors and about 30% of the subjunctive morphology. Italian exhibits greater dispersion across a broader array of governors, with *pensare* 'to think', and *credere* 'to believe', *sembrare* 'it seems', *bisognare* 'it is necessary' and *volere* 'to want' accounting for 44% of the governor pool and more than half of the subjunctive morphology in discourse.

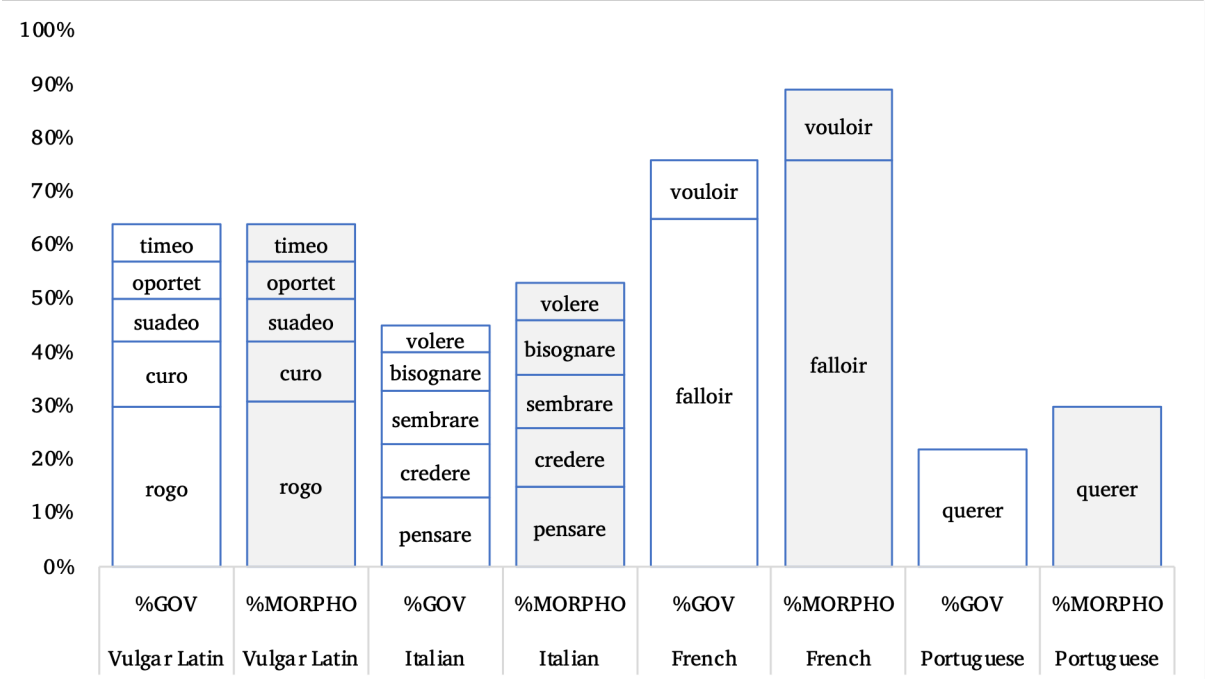


Figure 9: Dispersion of subjunctive morphology across governors by language, as measured by proportion frequent governors represent of the governor pool (%GOV) and proportion of subjunctive morphology they account for (%MORPHO), adapted from Poplack et al. (2018: 241).

Considering the semantic classes of governors and the identities of these in VL, there is evident transmission to the daughter languages, especially regarding volitive governors. This is exemplified by *falloir* 'it is necessary' (rate of subjunctive selection, 89%) and *vouloir* 'to want' (91%) in French, *querer* 'to want' (76%) in Portuguese, *bisognare* 'it is necessary' (100%) and *volere* 'to want' (92%) in Italian, indicating a preservation of these contexts within the Romance languages.

Additionally, Poplack and her associates (2018) highlighted that despite the eligibility of virtually all verbs for subjunctive morphology, only a limited number of

verbs actually carry it in discourse. This selective usage, responsible for a significant portion of subjunctive morphology in Portuguese, more than a third in both languages, and even more so in French (almost 70%) and Italian (where irregular forms of *essere* 'to be', mainly *sia*$_{[be.1\text{-}2\text{-}3SG.PRES.SUBJ]}$ contribute significantly to the selection of the subjunctive; Digesto 2019: 142), suggests a lexically-driven and non-productive use of the subjunctive in these languages.

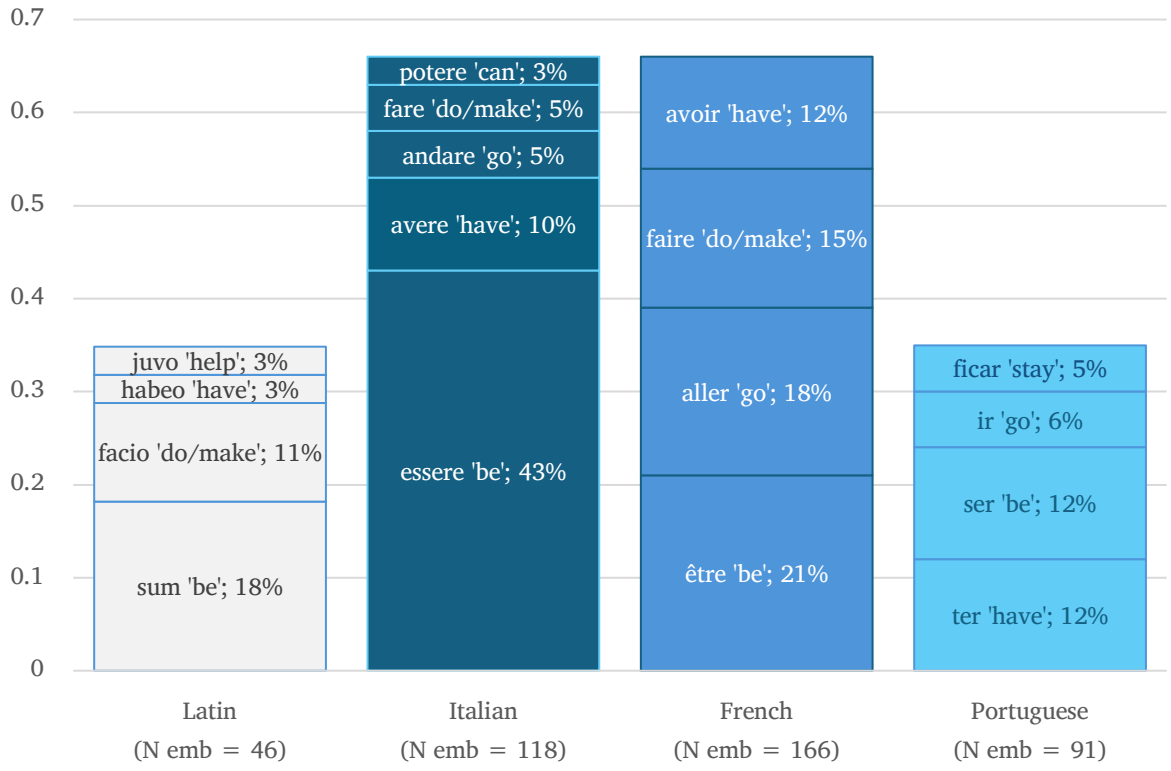


Figure 10: Distribution of subjunctive morphology across embedded verbs, adapted from Poplack et al. (2018: 243).

7. Discussion

This study contributes to our understanding of mood selection in completive clauses across Romance, offering an insight into the application of the variation method to examine typological relevant linguistic issues. We examined whether the lexicalized pattern of the subjunctive in contemporary Romance reflects a loss of semantic conditioning or the persistence of an inherited structure from Vulgar Latin. By analyzing VL data using established empirical methods, we traced these patterns to their origins and enabled a cross-linguistic comparison of subjunctive use in contexts

where these patterns remain operative. Despite the limitations of the data and the challenge of representing historical speech accurately, our findings reveal remarkable parallels between VL and Italian, extending to the broader Romance language domain, emphasizing a high "degrees of Romanceness" (Portner 1998: 38).

Our examination of the subjunctive in VL, particularly within Trimalchio's dinner party from the Satyricon, demonstrated a near-categorical selection of this mood in complement clauses, limited to a few governors. This observed pattern wasn't due to the VL corpus's size; a comparable 16th-century Italian corpus offered a greater diversity of governors and subjunctive governors. Furthermore, previous variationist research provided similar insights: the Canadian French corpus counts more than 2 million words; nevertheless, the corpus yielded an even smaller governor pool (n = 37!). This is related to the fact that the subjunctive is context-related, primarily influenced by the governor's lexical identity. However, almost all the VL governors belong to the semantic classes of volitive verbs. Surprisingly, the rate of subjunctive selection according to these semantic classes appeared fairly stable across time; governors within them highly favour or categorically favour the use of the subjunctive. While rates of subjunctive selection although decrease overtime (see *volere* 'to want' in both its affirmative and negative counterparts in VL, *volo* 'I want' and *nolo* 'I do not want'), these remain strong predictors of subjunctive selection, nonetheless accounting for less and less variation in discourse.

The prototypical use of the subjunctive in completive clauses could be considered 'harmonic' with the pre-existing paratactic constructions of Latin. This raises a critical question: is the previous paratactic (independent clause) meaning of the subjunctive still retained in the new hypotactic constructions, or is it just a syntactic mechanism devoid of any meaning and therefore reflects the grammaticalization of the Latin subjunctive as a subordinating marker? One perspective posits that the subjunctive was semantically motivated in VL, and given the transmission of its contexts of use to Romance languages (volitive verbs), we could consequently assume that these contexts have retained their original semantic contribution. Conversely, another interpretation challenges this semantic assumption, suggesting that the Latin subjunctive was already a morphosyntactic rather than a morphosemantic device, and that these morphosyntactic characteristics had been transferred into the daughter languages. The evidence at our disposal suggests a lack of lexical productivity in these so-called "semantic" contexts and a handful of governors responsible for the overall apparent semantic effect of these classes. Moreover, the surprising continuity of the

Latin pattern in Italian over the centuries, as well as French and Portuguese, weakens the semantic hypothesis due to the fact that these categories are both sparsely populated with governors, and infrequent in the data overall (see also Poplack et al. 2018). Evidence collectively points to a lexicalized nature of the use of the subjunctive in discourse. Furthermore, we can assume that the use of the subjunctive in VL was already "automatic", as previously suggested by Harris (1974), as evidenced by the categorical association of the subjunctive with a very small set of verbal governors, such as *rogo* 'to ask/to want', *oportet* 'it is necessary', among others.

In support of the choice of ruling out the question of the semantic role of volitive governors, there is also evidence that the overwhelming majority of the variation observed in Italian–both historically and in contemporary data–can be accounted for by the clear lexicalization of the subjunctive in discourse. As reported above, this stems from the fact that most of the variation in discourse is accounted by a handful of governors and by suppletive forms of *essere* 'to be'.

Based on the overall rates of subjunctive selection, the trajectory from VL to contemporary Romance languages shows a decline in the overall rate.[10] Despite the considerable temporal gap difference between the VL and the earlier Romance benchmark (16th century Italian), as well as the evolution within the Romance family, the analysis uncovered a widespread, consistent pattern across these languages. The subjunctive, though it has adapted to the lexical and syntactic characteristics of each language, maintains a core set of governing principles derived from VL. These play a crucial role in linguistic variation and account for most of the subjunctive morphology used in the targeted Romance varieties, notably the effect of *to want* across all three Romance languages, and *it is necessary* in Italian and more particularly in French. Nevertheless, an innovative trend has emerged: Italian, diverging from its Latin roots, has increasingly employed verbs of thinking such as *credere* 'to believe' and *pensare* 'to think' as subjunctive triggers. This development, where subjunctive forms are found in contexts not aligned with original Latin usage, is mirrored in other Romance languages, although they are quite rare in the variable context and, particularly in French Canadian speech, they highly disfavour the use of the subjunctive. Previous research into the dynamics of the lexical contribution in contemporary Italian subjunctive has identified a significant number of governors within what is classified

[10] In addition to the overall rate of subjunctive in contemporary Italian discourse reported in Figure 4 (66%), the use of the subjunctive, while quite robust in some Romance languages compared to others, shows a decline in French to 76% and more significantly in Portuguese to 55% (Poplack et al. 2018: 229).

as the semantic class of opinion verbs. Yet, the majority of variation and subjunctive use attributed to opinion verbs is concentrated amongst only a few governors, such as *credere* 'to believe', *pensare* 'to think', *sembrare* 'to seem'. Additional analyses investigating the supposed semantic contribution, examining the distinction between affirmative and negative constructions of verbs such as *credere* 'to believe' and *pensare* 'to think', along with the identification of contextual cues suggesting uncertainty or doubt, have not demonstrated a significant divergence in the choice between subjunctive and indicative forms under opinion verbs.

We observed a diminishing effect and a significant reduction in the distribution of the subjunctive with verbs of volition, which are the principal contexts of its use in VL data. Despite this, the daughter languages appear to still exhibit a perceived "semantic" impact of volitive verbs on subjunctive selection, though this effect appears to be actually an epiphenomenon resulting from the lexical identities characterizing these so-called semantic classes (see *querer* 'to want', *falloir* 'to be necessary', *vouloir* 'to want', *volere* 'to want', *bisognare* 'to be necessary'). The findings related to the contribution of a limited number of embedded verbs, which are predominantly responsible for subjunctive selection in Romance languages, further challenge the notion of a semantic contribution. Instead, they point to a lexicalized pattern of subjunctive use.

These findings lend support to the view that the idea of the subjunctive as semantically meaningful is a relatively modern development. For much of its documented history, the subjunctive was regarded primarily as a formal marker of subordination (*subiunctivus*, as its name underscores its hypotactic status) rather than as a carrier of inherent meaning. Metalinguistic analyses of normative grammars across centuries, both in Italian (Digesto 2019) and in French (Poplack & Dion 2013), similarly indicate that concerns about the meaning of the subjunctive emerge particularly in the modern period, coinciding with the rise of contemporary linguistic inquiry. Since the Latin era, the subjunctive was seen as morphosyntactically functional but semantically empty: *per se non exprimat sensum* ('by itself it does not express meaning,' Diomede, Latin grammarian; Keil ed. [1857: 340]).

8. Conclusion: bridging the past and future

The quantitative approach helped in elucidating grammatical (dis)similarities between languages beyond merely identifying the presence or absence of a form,

especially when syntactic patterns are expected to favour one form over another. Regarding the expected dissimilar syntactic outcomes between the ancestor language and Italian in particular, the subjunctive mood has broadened its subordinating role, a function that was already well-established by the 16th century in Italian, as exemplified by an increasing number of governors over time, even though a select few account for the majority of the variation in discourse.[11]

Furthermore, the shift from OV to VO structures in VL seems not to have impacted the lexical governors that trigger the subjunctive in the ancestor language. In fact, the findings indicate that the predominant strategy among governors is finite complementation and that the subjunctive mood has consistently served as the mood of subordination, primarily within completive clauses, a role that has remained relatively unchanged from VL to Italian. However, it's noteworthy that complement clauses featuring the subjunctive were initially triggered mainly by non-verbal governors, especially with *ut*, a pattern that has evolved in the descendant languages (see Figure 4: we observe increasing use of the subjunctive with verbal governors as opposed to non-verbal ones).

These results became apparent from a systematic and empirical examination of the subjunctive across languages and over time, adopting a corpus-based perspective to shed light on the processes governing its selection. Focusing on actual speech or close facsimiles of speech (such as dialogues written for stage plays) for historical benchmarks also allows for an exploration of inherent variability and its structured heterogeneity, and helps mitigate the potential influence of normative injunctions.

The systematic and exhaustive analysis of all instances where the subjunctive appears in the corpus, factoring in lexical, semantic, and syntactic factors, both in isolation and in combination, allows us to advance beyond mere overall occurrence rates. Relying solely on these rates, even within a defined semantic class, can obscure underlying patterns that aren't immediately apparent. Conversely, by

[11] It is worth noting that Digesto (2019; 2022) showed that the subjunctive is increasingly associated with suppletive forms of *essere* 'to be' and the complementizer *che* 'that' over time. Furthermore, in contemporary speech data, it has gained social prestige, evolving into a sociolinguistic marker: speakers with higher levels of education predominantly use the subjunctive within the variable context, employing a diverse array of governors alongside 'that' and suppletive forms of *essere*. However, no significant differences are observed in the use of the subjunctive with a core set of governors, such as *credere* 'to believe', *pensare* 'to think', *sembrare* 'to seem', *bisognare* 'to be necessary', and *volere* 'to want', across varying levels of speaker's education or speech styles.

clearly delineating a variable context, as well as by empirically operationalizing hypotheses about the contributions of linguistic (as well as extralinguistic) factors, we can discern cross-linguistic differences and similarities through quantifiable patterns of linguistic variation.

Abbreviations

1 = first person	F = feminine	PL = plural
2 = second person	IMPRF = imperfect tense	PRES = present tense
3 = third person	IND = indicative mood	SG = singular
ACT = active voice	M = masculine	SBJV = subjunctive mood

References

Acquaviva, Paolo. 1996. The Logical Form of Negative Concord. *Working Papers in Linguistics*, 1–29. Università Ca' Foscari di Venezia.

Ayres-Bennett, Wendy. 2000. Voices from the past. Sources of Seventeenth-Century Spoken French. *Romanische Forschungen,* 112(3). 323–348.

Berlinck, Rosane. 2019. Subjuntivo vs indicativo em orações completivas: percurso diacrônico no português brasileiro. In Ernestina Carrilho, Ana Maria Martins, Sandra Pereira & João Paulo Silvestre (eds.), E*studos linguísticos e filológicos oferecidos a Ivo Castro*, 217–245. 1st ed. Lisboa: Centro de Linguística da Universidade de Lisboa.

Binazzi, Neri. 2015. La frequente rinuncia al "che" nel parlato fiorentino: caratteristiche del fenomeno e spunti di riflessione per la lingua comune. *Studi di grammatica italiana*, vol. 33, 255–294. Le Lettere. Firenze: Accademia della Crusca.

Bodelot, Colette. 2003. *Grammaire fondamentale du latin: Les propositions complétives en latin.* Tome X (Bibliothèque d'études Classiques). Louvain: Peeters.

Bolkestein, A. Machtelt. 1989. Parameters in the expression of embedded predications in Latin. In Gualtiero Calboli (ed.), *Subordination and other Topics in Latin: Proceedings of the Third Colloquium on Latin Linguistics*, Bologna, 1-5 April 1985, 3–36. Amsterdam/Philadelphia: John Benjamins Publishing.

Bonomi, Ilaria. 1993. I giornali e l'italiano dell'uso medio. *Studi di grammatica italiana,* 15. 181-201.

Bybee, Joan L. 2003. Mechanisms of change in grammaticization: The role of frequency. In Brian Joseph & Richard Janda (eds.), *The Handbook of Historical Linguistics*, 602–623. London: Blackwell.

Bybee, Joan L., Revere Perkins & William Pagliuca. 1994. *The Evolution of Grammar: Tense, Aspect and Modality in the Languages of the World.* Chicago: University of Chicago Press.

Campbell, Lyle. 2013. *Historical Linguistics: An Introduction*. 3rd ed. Cambridge/Massachusetts: The MIT Press.

Carlier, Anne, Walter De Mulder & Béatrice Lamiroy. 2012. Introduction: The pace of grammaticalization in a typological perspective. *Folia Linguistica,* 46(2). 287–302. (doi:10.1515/flin.2012.010)

Chinellato, Paolo. 2001. L'interpretazione morfosemantica del modo congiuntivo in italiano e in tedesco. *Rivista di grammatica generativa* 26. 3–20.

Clackson, James. 2011. *A Companion to the Latin Language.* Chichester, UK: John Wiley & Sons.

Costantini, Francesco. 2011. Subjunctive obviation in nonargument clauses. *Working Papers in Linguistics,* 39–61. Università Ca'Foscari di Venezia.

Cressot, Marcel. 1947. *Le style et ses techniques.* Paris: Presses Universitaires de France.

Cresti, Emanuela & Massimo Moneglia. 2005. *C-ORAL-ROM integrated reference corpora for spoken Romance languages.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

(http://public.eblib.com/choice/publicfullrecord.aspx?p=622285) (Accessed May 30, 2016.)

De Mauro, Tullio, Federico Mancini, Massimo Vedovelli & Miriam Voghera. 1993. *Lessico di frequenza dell'italiano parlato.* Milano: Etas.

Digesto, Salvio. 2019. Verum a fontibus haurire. *A Variationist Analysis of Subjunctive Variability Across Space and Time: from Contemporary Italian back to Latin.* Ottawa, Canada: University of Ottawa.

Digesto, Salvio. 2021. Lexicalization and Social Meaning of the Italian Subjunctive. *Cadernos de Linguística*, 2(3). e609. (doi:10.25189/2675-4916.2021.V2.N3.ID609)

Disterheft, Dorothy & Carlotta Viti. 2010. Subordination. In Silvia Luraghi & Vit Bubenik (eds.), *Continuum companion to historical linguistics* (Bloomsbury Companions), 230–249. London: Continuum.

Dorigo, Carlo Alberto. 1951. Problemi di metodo e prospettive nuove negli studi della sintassi italiana. *Aevum. Rassegna di scienze storiche, linguistiche e filologiche*, XXV(4). 355–361.

Durante, Marcello. 1981. *Dal latino all'italiano moderno: saggio di storia linguistica e culturale.* Bologna: Zanichelli.

Elsig, Martin & Shana Poplack. 2006. Transplanted dialects and language change: question formation in Québec. *Penn Working Papers in Linguistics*, 12(2), Selected Papers from NWAV 34. 77–90.

Farkas, Donka F. 1992. On the semantics of subjunctive complements. In Paul Hirschbühler & Ernst Friedrich K. Koerner (Eds.), *Romance Languages and Modern Linguistic Theory* (Current Issues in Linguistic Theory, vol. 91), 69–104. Amsterdam: John Benjamins.

Gatta, Francesca. 2002. Osservazioni sul congiuntivo in margine ad una giornata televisiva. In Leo Schena, Michele Prandi & Marco Mazzoleni (eds.), *Intorno al congiuntivo, Atti del Convegno Nazionale "Attorno al congiuntivo. Storia, tipologia, traduzione" (Forlì, 2 – 3 marzo 2000)*, 83–92. CLUEB. Bologna: CLUEB.

Giannakidou, Anastasia. 1995. Subjunctive, habituality and negative polarity. *Semantics and Linguistic Theory* (SALT), 5. 132–150.

Giannakidou, Anastasia & Alda Mari. 2015. Mixed (Non)Veridicality and Mood Choice in Complement Clauses. *CLS* 51. Chicago, France.

Giorgi, Alessandra & Fabio Pianesi. 1997. *Tense and Aspect: from Semantics to Morphosyntax (Oxford Studies in Comparative Syntax)*. New York: Oxford University Press.

Givón, Talmy. 2018. *Negation in language: Between semantics and pragmatics.* On Understanding Grammar: Revised Edition. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Gorman, Kyle & Daniel Ezra Johnson. 2013. 214 Quantitative Analysis. In Robert Bayley, Richard Cameron, & Ceil Lucas (eds.), *The Oxford Handbook of Sociolinguistics,* 214–240. Oxford: Oxford University Press.

Guy, Gregory R. 1988. Advanced VARBRUL analysis. In Kathleen Ferrara, Becky Brown, Keith Walters & John Baugh (eds.), Linguistic Contact and Change: Proceedings of the Sixteenth Annual Conference on New Ways of Analyzing Variation, 124–136. Austin, TX: University of Texas Department of Linguistics.

Haiman, John. 1994. Ritualization and the development of language. In William Pagliuca (ed.), *Perspective on Grammaticalization*, 3–28. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Handford, Stanley A. 1947. *The Latin Subjunctive. Its Usage And Development From Plautus To Tacitus.* London: Methuen.

Harris, Martin. 1974. The subjunctive mood as a changing category in Romance. In John M. Anderson & Charles Jones (eds.), *Historical Linguistics II*, 169–188. Amsterdam: North-Holland Publishing Company.

Harris, Martin. 1978. *The evolution of French syntax: a comparative approach* (Longman Linguistics Library 22). London/New York: Longman.

Harris, Martin & Nigel Vincent. 1997. *The Romance Languages*. London: Routledge.

Haspelmath, Martin. 2003. The geometry of grammatical meaning: semantic maps and cross-linguistic comparison. In Michael Tomasello (ed.), *The new psychology of language: cognitive and functional approaches to language structure,* vol. 2, 211–242. Mahwah, New Jersey: Lawrence Erlbaum.

Haudry, Jean. 1973. Parataxe, hypotaxe et correlation dans la phrase latine. *Bulletin de la Société de Linguistique de Paris*, 68. 147–186.

Herman, József. 1989. Accusativus cum Infinitivo et subordonnée à quod, quia en latin tardif: Nouvelles remarques sur un vieux problème. In Gualtiero Calboli (ed.), *Subordination and Other Topics in Latin: Proceedings of the Third Colloquium on Latin Linguistics*, Bologna, 1-5 April 1985, 133–152. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Hooper, Joan. 1975. On assertive predicates. In John Kimball (ed.), *Syntax and semantics*, vol. 4, 91–124. New York: Academic Press.

Jespersen, Otto. 1924. *The philosophy of grammar.* London: Allen & Unwin.

Kastronic, Laura. 2016. *A Comparative Variationist Approach to Morphosyntactic Variation in Contemporary Hexagonal and Quebec French.* Ottawa, Canada: University of Ottawa.

Keil, Heinrich. 1857. *Flavii Sospatri Charisii Artis grammaticae libri v. Diomedis Artis grammaticae libri III. Ex Charisii Arte grammatica excerpta*. Lipsia: In aedibvs B.G. Teubneri.

Klein, Flora. 1975. Pragmatic constraints on distribution: the Spanish subjunctive. *Papers from the regional meeting of the Chicago Linguistic Society* XI. 353–365.

Labov, William. 1994. *Principles of Linguistic Change, Internal Factors. Language in Society*. Vol. 1: Internal Factors. Cambridge and Oxford: Blackwell Publishers.

Labov, William. 1972. *Sociolinguistic Patterns.* Philadelphia: University of Pennsylvania Press.

Labov, William. 1984. Field Methods of the Project on Linguistic Change and Variation. In John Baugh & Joel Sherzer (eds.), *Language in Use: Readings in Sociolinguistics*, 28–54. Englewood Cliffs, NJ: Pretence-Hall. (Accessed July 17, 2017.)

Labov, William. 1989. The child as linguistic historian. *Language Variation and Change*, 1(1). 85–97.

Labov, William. 2001. *Principles of Linguistic Change, Social Factors. Language in Society*. Vol. 2: Social Factors. Malden and Oxford: Blackwell Publishers.

Lehmann, Christian. 1995. *Thoughts on Grammaticalization*. 2nd revised. Munich: Lincom.

Leone, Alfonso. 1949. *L'uso del congiuntivo in Latino. Principi di sintassi ragionata.* Catania: tip. Coniglione & Giuffrida.

Lindschouw, Jan. 2011. *Étude des modes dans le système concessif en français du 16e au 20e siècle et en espagnol moderne : évolution, assertion et grammaticalisation.* Copenhagen: Museum Tusculanum Press.

Loengarov, Alexander. 2006. L'alternance indicatif/subjonctif dans les langues romanes. Motivation sémantico-pragmatique et grammaticalisation. Doctoral dissertation, University of Leuven (KUL).

Lombardi Vallauri, Edoardo. 2003. Pragmaticizzazione dell'incompletezza sintattica nell'italiano parlato: le ipotetiche sospese. In *Il parlato italiano*.

Magni, Elisabetta. 2009. Mood and Modality. In Philip Baldi & Pierluigi Cuzzolin (eds.), *New Perspectives on Historical Latin Syntax. Constituent Syntax: Adverbial Phrases, Adverbs, Mood, Tense*, Vol. 2, 193–275. Berlin/New York: Mouton de Gruyter.

Maiden, Martin & Cecilia Robustelli. 2007. *A Reference Grammar of Modern Italian* (Routledge Reference Grammars). 2nd edn. London/New York: Routledge.

Manzini, Maria Rita. 2000. Sentential Complementation: The subjunctive. In Peter Coopmans, Martin Everaert & Jane Grimshaw (eds.), *Lexical Specification and Insertion,* 241–267. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Mellet, Sylvie. 1994. Le subjonctif. In Sylvie Mellet, Marie-Dominique Joffre & Guy Serbat (eds.), *Grammaire fondamentale du latin. Le signifié du verbe*, 173–209. Louvain: Peeters.

Murphy, Melissa Dae. 2008. The role of typological drift in the development of the Romance subjunctive: a study in word-order change, grammaticalization and synthesis. Doctoral dissertation: The University of Texas, Austin.

Nevalainen, Terttu & Helena Raumolin-Brunberg. 2017. *Historical Sociolinguistics: Language Change in Tudor and Stuart England.* 2nd Edition. London: Routledge.

Noonan, Michael. 2007. Complementation. In Timothy Shopen (ed.), *Language Typology and Syntactic Description,* Vol. 2, 52–150. Complex Constructions. Cambridge: Cambridge University Press.

Palmer, Leonard R. 1988. *The Latin Language (The Great Languages)*. Norman: University of Oklahoma Press.

Palmer, Frank R. 2001. *Mood and Modality*. Cambridge University Press.

Perotti, Pier Angelo. 1996. Sulle interrogative indirette in latino. *Latomus: revue d'études latines,* 55(3–4). 329–338.

Pinkster, Harm. 1990. *Latin syntax and semantics.* London: Routledge.

Poletto, Cecilia. 2000. *The Higher Functional Field: Evidence from the Northern Italian Dialects.* New York: Oxford University Press.

Poplack, Shana. 1989. The care and handling of a mega-corpus: The Ottawa-Hull French project. In Ralph W. Fasold & Deborah Schiffrin (eds.), *Language Change and Variation* (Current Issues in Linguistic Theory 52), 411–451. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Poplack, Shana. 1990. Prescription, intuition et usage: Le Subjonctif français et la variabilité inhérente. *Langage et Societe,* 54. 5–33. (doi:10.3406/lsoc.1990.2499)

Poplack, Shana. 1992. The inherent variability of the French subjunctive. In Christiane Laeufer & Terrell A. Morgan (eds.), *Theoretical analyses in Romance linguistics: selected papers from the nineteenth Linguistic Symposium on Romance Languages* (LSRL XIX), The Ohio State University, April 21-23, 1989, 235–263. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Poplack, Shana. 2011. Grammaticalization and linguistic Variation. In Bernd Heine & Heiko Narrog (eds.), *The Oxford Handbook of Grammaticalization,* 209–224. Oxford: Oxford University Press.

Poplack, Shana, Allison Lealess & Nathalie Dion. 2013. The evolving grammar of the French subjunctive. *International Journal of Latin and Romance Linguistics*, 25(1). 139–195. (doi:10.1515/probus-2013-0005)

Poplack, Shana & Stephen Levey. 2010. Contact-induced grammatical change: A cautionary tale. In Peter Auer & Jürgen Erich Schmidt (eds.), L*anguage and Space - An international handbook of linguistic variation*, Vol. 1: Theories and methods, 391–419. Berlin: Mouton de Gruyter.

Poplack, Shana & Sali A Tagliamonte. 2001. *African American English in the Diaspora.* Oxford: Basil Blackwell.

Poplack, Shana, Rena Torres Cacoullos, Nathalie Dion, Rosane e Andrade Berlinck, Salvio Digesto, Dora Lacasse & Jonathan Steuck. 2018. Variation and grammaticalization in Romance: A cross-linguistic study of the subjunctive. In Wendy Ayres-Bennett & Janice Carruthers (eds.), *Manuals in Linguistics: Romance Sociolinguistics*, 217-252. Berlin: Mouton de Gruyter.

Portner, Paul. 1998. The semantics of mood. *GLOT International*, 4(1). 3-8

Portner, Paul & Aynat Rubinstein. 2013. Mood and contextual commitment. In *Proceedings of SALT, 22*. 461–487. CLC Publications.

Quer, Josep. 2001. Interpreting mood. Probus. *Walter de Gruyter,* 13(1). 81–111. (doi:10.1515/prbs.13.1.81)

Romaine, Suzanne. 1982. *Socio-historical linguistics: its status and methodology*. Cambridge: Cambridge University Press.

Santulli, Francesca. 2009. Il congiuntivo italiano: morte o rinascita? In *Rivista Italiana di Linguistica e di Dialettologia*. 167–195.

Scarsi, Mariangela. 1996. *Satyricon / Petronius* (Classici Giunti). Firenze: Giunti.

Schachter, Paul & Timothy Shopen. 2007. Parts-of-speech systems. In Timothy Shopen (ed.), *Language Typology and Syntactic Description*. Vol. 1: Clause Structure, 1–60. 2nd edn. Cambridge: Cambridge University Press. (doi:10.1017/CBO9780511619427.001)

Schmitt Jensen, Jørgen. 1970. *Subjonctif et hypotaxe en italien: une esquisse de la syntaxe du subjonctif dans les propositions subordonnées en italien contemporain.* Odense: Odense University Press.

Schneider, Stefan. 1999. *Il congiuntivo tra modalità e subordinazione: uno studio sull'italiano parlato*. Roma: Carocci.

Shlonsky, Ur. 2006. Projection étendue et cartographie de SC. *Nouveaux cahiers de linguistique française,* (27). 83–93.

Simone, Raffaele. 1993. Stabilità e instabilità nei caratteri originali dell'italiano. In Alberto Sobrero & Paola Benincà & Gaetano Berruto (eds.), *Introduzione all'italiano contemporaneo*. Vol. I: Le strutture. 41–100. Roma - Bari: Laterza.

Soliman, Luciana T. 2002. *Modalità e implicazioni aspettuali: analisi contrastiva italiano/francese del congiuntivo nell'ottica psicomeccanica.* In Atti del Convegno Attorno al congiuntivo. Storia, tipologia, traduzione. 291–306.

Stefanelli, Stefania. 2006. *Va in scena l'italiano: la lingua del teatro tra Ottocento e Novecento.* Firenze: Cesati.

Terrell, Tracy & Joan Hooper. 1974. A semantically based analysis of mood in Spanish. *Hispania,* 57. 484–494.

Thompson, Sandra A. 1998. A discourse explanation for the cross-linguistic differences in the grammar of incorporation and negation. In Anna Siewierska & Ja Jung Song (eds.), *Case, Typology and Grammar*. Vol. 38, 309–342. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Veland, Reidar. 1991. L'influence modale des verbes d'opinion en italien contemporain: Étude d'un corpus. *Studia Neophilologica: A Journal of Germanic and Romance Languages and Literature*, 63(2). 209–220.

Voghera, Miriam. 1993. La grammatica del LIP. In Tullio De Mauro, Federico Mancini, Massimo Vedovelli & Miriam Voghera (eds.), *Lessico di frequenza dell'italiano parlato*, 86–111. Milano: Etas.

Voghera, Miriam. 2001. Teorie linguistiche e dati di parlato. In Francesco Albano Leoni, Lucia Cresti & Giorgio Fiorentino (eds), *Dati empirici e teorie linguistiche*, 75-95. Roma: Bulzoni.

Wandruszka, Ulrich. 1998. Frasi subordinate al congiuntivo. In Lorenzo Renzi & Giampaolo Salvi (eds.), *Grande grammatica italiana di consultazione*, vol. II, 415–481. Bologna: Il Mulino.

Whaley, Lindsay. 1997. *Introduction to Typology: The Unity and Diversity of Language.* Thousand Oaks, CA: SAGE Publications. (doi:10.4135/9781452233437)

Appendix

This appendix lists all matrix governors that selected the subjunctive at least once in the contemporary Italian dataset. For each, it provides an English translation, the subjunctive rate, token count, proportion of overall data (% Data), proportion of total subjunctive forms (% Morpho), and the increment, a metric indicating whether the

governor contributes disproportionately more subjunctive morphology than expected given its frequency in the data.

Governor	Translation	% Subj	N	% Data	% Morpho	Increment
non è	it is not	33%	37/112	18%	9%	-9%
(non) pensare	to (not) think	73%	64/88	14%	16%	2%
(non) sembrare	to (not) seem	75%	48/64	10%	12%	1%
(non) credere	to (not) believe	79%	45/57	9%	11%	2%
bisognare	to be necessary	93%	41/44	7%	10%	3%
(non) parere	to (not) appear	58%	15/26	4%	4%	-1%
(non) dire	to (not) say	19%	5/26	4%	1%	-3%
(non) volere	to (not) want	84%	16/19	3%	4%	1%
può darsi	it might be	58%	11/19	3%	3%	0%
sperare	to hope	93%	14/15	2%	3%	1%
bastare	to be enough / suffice	50%	6/12	2%	1%	0%
(non) essere sicuro	to (not) be sure	63%	5/8	1%	1%	0%
aspettare	to wait / to expect	50%	4/8	1%	1%	0%
non sapere	to not know	33%	2/6	1%	0%	0%
è facile	it is easy	100%	4/4	1%	1%	0%
è importante	it is important	100%	4/4	1%	1%	0%
avere paura	to be afraid	75%	3/4	1%	1%	0%
è meglio	it is better	75%	3/4	1%	1%	0%
fare sì	to ensure / to make sure	75%	3/4	1%	1%	0%
avere l'impressione	to have the impression	50%	2/4	1%	0%	0%
è inutile	it is useless	50%	2/4	1%	0%	0%
immaginare	to imagine	25%	1/4	1%	0%	0%
(non) è bene	to (not) be good	100%	3/3	0%	1%	0%
avere la sensazione	to have the feeling	100%	3/3	0%	1%	0%
esigere	to demand	100%	3/3	0%	1%	0%
non è detto	to not be certain	100%	3/3	0%	1%	0%
ritenere	to consider	100%	3/3	0%	1%	0%
supporre	to suppose	100%	3/3	0%	1%	0%
presupporre	to presuppose	67%	2/3	0%	0%	0%
richiedere	to require / to demand	67%	2/3	0%	0%	0%
(non) è possibile	to be (not) possible	33%	1/3	0%	0%	0%
è ovvio	it is obvious	33%	1/3	0%	0%	0%
ci sta	it is possible	100%	2/2	0%	0%	0%
è difficile	it is difficult	100%	2/2	0%	0%	0%
è impossibile	it is impossible	100%	2/2	0%	0%	0%
mettere	to assume	100%	2/2	0%	0%	0%
non succedere	to not happen	100%	2/2	0%	0%	0%

Governor	*Translation*	*% Subj*	*N*	*% Data*	*% Morpho*	*Increment*
può essere	it may be	100%	2/2	0%	0%	0%
calcolare	to calculate	50%	1/2	0%	0%	0%
dubitare	to doubt	50%	1/2	0%	0%	0%
fare in maniera	to ensure	50%	1/2	0%	0%	0%
non è vero	to not be true	50%	1/2	0%	0%	0%
trovare	to find	50%	1/2	0%	0%	0%
(non) dispiacersi	to (not) mind	100%	1/1	0%	0%	0%
(non) è giusto	to (not) be fair	100%	1/1	0%	0%	0%
assicurarsi	to make sure	100%	1/1	0%	0%	0%
avere il dubbio	to have doubt	100%	1/1	0%	0%	0%
controllare	to check	100%	1/1	0%	0%	0%
dedurre	to deduce	100%	1/1	0%	0%	0%
è bello	it is nice	100%	1/1	0%	0%	0%
è inevitabile	it is inevitable	100%	1/1	0%	0%	0%
è strano	it is strange	100%	1/1	0%	0%	0%
essere orgoglioso	to be proud	100%	1/1	0%	0%	0%
evitare	to avoid	100%	1/1	0%	0%	0%
fare in modo	to ensure / to make sure	100%	1/1	0%	0%	0%
fare piacere	to please	100%	1/1	0%	0%	0%
l'importante è	the important thing is	100%	1/1	0%	0%	0%
lasciare	to let / to allow	100%	1/1	0%	0%	0%
non avere il dubbio	to have no doubt	100%	1/1	0%	0%	0%
non avere senso	to make no sense	100%	1/1	0%	0%	0%
non è da dire	it is not to be said	100%	1/1	0%	0%	0%
non è umano	it is not humane	100%	1/1	0%	0%	0%
non preoccuparsi	to not worry	100%	1/1	0%	0%	0%
preferire	to prefer	100%	1/1	0%	0%	0%
presumere	to presume	100%	1/1	0%	0%	0%
reputare	to deem / to consider	100%	1/1	0%	0%	0%
rischiare	to risk	100%	1/1	0%	0%	0%
sentire il timore	to feel fear	100%	1/1	0%	0%	0%
servire	to be needed	100%	1/1	0%	0%	0%
sorprendere	to surprise	100%	1/1	0%	0%	0%
verificare	to verify	100%	1/1	0%	0%	0%
Total		**66%**	**404/616**			

CONTACT

salvio.digesto@carleton.ca

Linguistic Typology at the Crossroads 5-2 (2025): 289-322 ISSN 2785-0943

Object encoding in spoken language data and antipassives

SILVIA BALLARÈ[1], CATERINA MAURI[1], ANDREA SANSÒ[2]
[1]ALMA MATER STUDIORUM - UNIVERSITY OF BOLOGNA
[2]UNIVERSITY OF SALERNO

Submitted: 15/04/2025 Revised version: 12/01/2026
Accepted: 18/01/2026 Published: 25/02/2026

Abstract

This paper investigates object deletion in spoken Italian as a strategy for backgrounding the patient argument, drawing a parallel with antipassive (AP) constructions in a cross-linguistic perspective. Although Italian lacks a dedicated morphological AP, we examine whether the semantic and discourse-related factors known to condition AP derivation, such as agent affectedness, telicity, and contextual accessibility, also shape patterns of object realization in spontaneous discourse. Based on a corpus-driven analysis of 1.500 occurrences of 15 transitive verbs from the KIParla corpus, we combine a conditional inference tree and a random forest with a mixed-effects logistic regression model to assess the relative contribution of semantic, syntactic, and discourse-pragmatic predictors, while controlling for lexical variability. The results show that object deletion in Italian is primarily governed by contextual accessibility, which emerges as the strongest and most robust predictor across modelling techniques. Agent affectedness further modulates realization preferences, facilitating object deletion in event types that foreground changes in or consequences for the agent. By contrast, the other predictors do not exert independent effects once accessibility and lexical differences are taken into account. Overall, the findings highlight the importance of spoken data for a discourse-sensitive typology of argument realization.

Keywords: antipassive; object deletion; spoken data; Italian; contextual accessibility.

1. Introduction: definition, research questions and objectives

Antipassives (APs)[1] are a type of grammatical construction found in many languages, characterized by the demotion or omission of the patient of a transitive clause. There are two primary definitions of APs, one narrow and one broad. Under a narrow definition (e.g. Dixon & Aikhenvald 2000: 9), in APs the patient is either omitted or marked as a non-core argument, using a different case or an adpositional phrase, and explicit morphological marking on the verb is required. This results in a derived intransitive clause where the verb remains linked to the original transitive meaning but with reduced or no emphasis on the patient. Consider (1):

(1) Chukchi (ckt; Chukotko-Kamchatkan, Chukotian; Dunn 1999: 200)

a. *ʔaatcek-a* *piri-nin* *roolqəl*
youth-ERG take-3SG.A > 3SG.OBJ food.3SG.ABS
'The youth took the food.' (transitive)

b. *ʔaatcek* ***ine****-piri-ɣʔi*
youth.3SG.ABS AP-take-TH.3SG.SBJ
'The youth took (something) / the youth won the prize. (idiomatic interpretation)' (AP)

In (1a), the patient appears in the absolutive form, which is used in Chukchi to mark objects in transitive constructions, while the agent takes the ergative suffix *-a*. In (1b), the patient is omitted, and the sole remaining argument, the agent, is in the absolutive case, as is typical for intransitive subjects. Moreover, the verb is marked with the AP prefix *ine-*.

Broader definitions (e.g., Polinsky 2013) also encompass constructions in which the patient is demoted to a non-core argument or a non-argument position, regardless of whether the verb carries any morphological marking. In Kalkatungu transitive constructions, the agent and the patient appear in the ergative and absolutive case, respectively, as exemplified in (2a). Another possibility is that the agent appears in the absolutive case, while the patient is either demoted to the dative case or deleted,

[1] The paper is the result of a continuous collaboration between the authors. Andrea Sansò is responsible for Section 1, Silvia Ballarè is responsible for Section 2 and Caterina Mauri is responsible for section 3, while the three authors are jointly responsible for Section 4.

with the verb remaining unmarked, as shown in (2b). The construction in (2b) can be considered as an AP construction under a broader definition of AP:

(2) Kalkatungu (ktg; Pama-Nyungan, Kalkatungic; Blake 1982: 86)

a. *ṭuka-yu* *tuar* *ityayi*
dog-ERG snake.ABS bite
'The dog bites/bit the snake.'
b. *ṭuku* *tuar-ku* *ityayi*
dog.ABS snake-DAT bite
'The dog is biting the snake.'

Another controversial issue concerns the existence of APs in non-ergative languages. In accusative languages, constructions that superficially resemble those in (2) can indeed be found; however, in such languages the demotion or omission of the patient does not result in a change in the flagging of the sole remaining argument, namely the agent (Janic 2021: 261).

(3) Italian (ita; Indo-European, Romance; personal knowledge)

a. *Marco* *ha* *mangiato* *gli* *spaghetti*
Marco AUX eat:PST.PTCP DEF.PL spaghetti:PL
'Marco ate spaghetti.'
b. *Marco* *ha* *mangiato*
Marco AUX eat:PST.PTCP
'Marco ate.'

A construction like (3b) is typically not labelled as AP in grammatical descriptions; instead, it is discussed as a case of object deletion construction. However, both APs and object deletion constructions tend to occur in the same or very similar pragmatic and discourse contexts—for example, to deemphasize the patient, highlight the action or agent, or signal that the object is irrelevant or non-specific. As Cooreman (1994: 51) puts it, both types of constructions reflect "a certain degree of difficulty in recognizing an effect resulting from an activity by A[gent] on an identifiable

O[bject]". In many languages, in fact, object deletion constructions constitute the only available strategy for backgrounding the patient.

In nearly half of the languages that exhibit APs, this construction is lexically restricted, meaning APs can only occur with a subset of transitive verbs (Heaton 2017; Polinsky 2017; Sansò 2017, 2018; Say 2021). For example, Heaton (2017) found that out of 134 languages with APs, 46 had fully productive APs (34,33%), 57 had partially productive APs (42,54%), and 31 had unproductive APs (23,13%). Polinsky (2013) reported similar rates of non-productive or partially productive APs in her analysis of WALS data, with around 40% falling into these categories. This raises an important question: do the lexical restrictions on APs correlate with similar tendencies characterizing object deletion constructions in accusative languages?

To address this question, this article explores the hypothesis that the lexical restrictions observed in APs across languages correlate with patterns of object deletion in spoken Italian. Specifically, we will examine (a) whether the syntactic, semantic, and pragmatic factors identified in the typological literature as relevant to AP derivation also influence object deletion in spoken Italian, and (b) whether these factors are equally important or can be ranked according to their influence. To understand the relationship between object deletion constructions and APs, it is crucial to consider tendencies observed in spoken language. Many typological studies on APs rely on elicited or non-spontaneous data, which may not fully capture the patterns emerging in natural discourse. Spontaneous speech may reveal grammatical tendencies that may remain unnoticed in controlled environments, particularly regarding constructions influenced by discourse structure and interaction. If object deletion follows patterns similar to those found in APs across languages, investigating spoken data could shed light on the functional motivations of these tendencies, refining our understanding of APs across languages.

As already discussed, Italian lacks a productive AP construction in the strict sense, and the most common strategy used when the patient is irrelevant, non-specific, or must be omitted for any other reasons connected to pragmatics of information structure is object deletion, making it an intriguing case for exploring the intersection of object deletion and APs. However, before analyzing spoken Italian data, it is essential to summarize what is known about lexical restrictions on APs across languages and to explain why Italian was chosen as a test case for this investigation. These tasks will be addressed in the following sections.

1.1. The productivity of APs: lexical and syntactic restrictions

One of the key findings in typological studies of APs is that their productivity can vary significantly across languages. As mentioned earlier, only 34,33% of the APs in Polinsky's (2013) sample are fully productive. For example, in Tolowa, there is a non-productive AP limited to three verbs. The meanings of the derived APs in (4a) and (4b) are somewhat lexicalized, whereas the verb meaning 'to eat' maintains the same meaning when the AP prefix *ch'ee-* is present:

(4) Tolowa (tol; Athabaskan-Eyak-Tlingit, Athabaskan; Givón & Bommelyn 2000: 53-54)

a.	*'u-sh-d-t'e'sr*	>	*ch'ee-d-t'e'sr*
	TH-1SG-D-tattoo		AP-D-tattoo
	'I am tattooing him/her.'		'S/he is writing.'
b.	*'u-sh-shush*	>	*ch'ee-shush*
	TH-1SG-sip		AP-sip
	'I am sipping it'.		'S/he is drinking alcohol.'
c.	*ch'ee-sh-ą*		
	AP-1SG-eat		
	'I am eating.'		

In Bezhta (kap; Nakh-Daghestanian, Tsezic; Khalilova & Comrie 2013), only 4 verbs out of the 83 verbs included in the ValPal database regularly participate in the AP/ergative alternation. As an example of a language with a productive AP, we can look at Kuku Yalanji, where the AP "is used productively for what may be called a 'generalised action' […], not discrete and […] performed on some general or 'non-individuated' object" (Patz 2002: 152):

(5) Kuku Yalanji (gvn; Pama-Nyungan, Yimidhirr-Yalanjic-Yidinic; Patz 2002: 153)

a.	*yinya*	*karrkay*	*kaya-nda*	*kuni-n-kuni-ji-y*
	that.ABS	child.ABS	dog-LOC.PT	hit-N-RED-AP-NPST
	'That little one is hitting all the dogs (around here).'			

b. *barna* *dunga-ny* *bunjurri-ji-ny*
Aborigine.ABS go-PST throw.spit/curse-AP-PST
'The Aborigine went and threw curses everywhere.'

c. *jalbu* *wukay-rnba* *nubi-nubi-ji-y*
woman.ABS yam-LOC search.for-RED-AP-NPST
'The woman is looking around for yam.'

The lexical restrictions on APs do not appear to be random. Heaton (2017) observed that in several languages, APs occur only with certain verb types, particularly those that are not highly transitive. Sapién et al. (2021) analyzed natural data from Cariban languages and identified three event types in which APs are more likely to appear: (i) routine processes where the agent performs an activity with a predictable and relatively insignificant patient (e.g. *eat*, *cut*, *paddle*, *cook*, etc.), (ii) cognitive events where the agent is the experiencer and the patient is the stimulus (e.g. *hear*, *understand*, etc.), and (iii) speech events where the agent produces speech and the patient represents what is said (e.g. *command*, *ask*, *tell*, etc.). In these cases, the missing patient can even be specific and identifiable (e.g., visible in the stimulus video), but it is simply not encoded in the verbal description (Sapién et al. 2021: 82).

The most comprehensive effort to identify the semantic properties of verbs that more readily favour AP derivation is by Say (2021). Say (2021) argues that the following semantic properties are strongly correlated with a verb's likelihood of appearing in the AP construction:

a) Agentivity of the agent: The subject is typically animate, volitional, and in control of the action.
b) Specification of the agent's manner: Some APs are associated with verbs that lexically encode the manner of action (e.g., *thresh* vs. *beat*).
c) Inherent atelicity: APs are more often found with verbs where telicity is determined by the incremental nature of the object (e.g., *eat* vs. *kill*).
d) Narrow class of potential patients: APs tend to occur more frequently with verbs that semantically select a limited class of patients (e.g., *eat* vs. *create*).
e) Affectedness of the agent: In some APs, the agent is not only the causer but also the endpoint, experiencing a significant change in state (e.g., *eat, drink*, *hear*).

Say (2021: 185) considers properties a)-e) as "contributing factors that determine the degree to which an individual verb is likely to behave as a natural antipassive." Natural APs are intended as a fuzzy category comprising verbs that are more likely to allow for AP derivation, similar to how verbs like *kiss* and *meet* can be considered natural reciprocals.

Another factor influencing the distribution of APs is finiteness. Some languages allow APs only with non-finite verb forms (e.g., Ch'ol, see Coon 2013; Mandinka, see Creissels 2012, 2015). Consider the following example from Mandinka:

(6) Mandinka (mnk; Mande, Western Mande; Creissels 2015: 240)

a. *ŋ* *ŋá* *mus-óo* *tuu-ri-tôo* *jé*
1SG COMP.POS woman-DEF pound-AP-SIMULT see
'I saw the woman pounding.'

b. *mus-óo* *be* *tuu-r-óo* *la*
woman-DEF COP pound-AP-DEF OBL
'The woman is pounding.' (lit: 'The woman is at the pounding.')

The AP formative *-ri* in Mandinka is used only when the verb is in a non-finite form that expresses temporal simultaneity, as in (6a), or when it is used nominally, as in (6b).

Do the factors identified by Say (2021) and the finiteness of the verb also affect the likelihood of object deletion? We will address this question by analyzing naturally occurring data from a sample of spoken Italian, a language that lacks a morphological AP. Before we delve into this analysis, it is important to provide a brief overview of object deletion in Italian, which will be covered in the next section.

1.2. Object deletion in Italian

In Italian, only a small set of verbs exhibit an AP marked by the reflexive clitic *si*. These include pairs such as *ricordare* vs. *ricordarsi* 'remember', *lamentare* vs. *lamentarsi* 'complain', and *vantare* vs. *vantarsi* 'boast'. The reflexive-marked AP permits both object deletion, as illustrated in (7b), and object demotion, as in (7c), where the patient is realized as an oblique argument in a *di*-headed prepositional phrase:

(7) Italian

a. *mi* *scusi* *mi* *scusi* *le* *sto*
1SG.OBJ excuse:SBJV.2SG 1SG.OBJ excuse:SBJV.2SG to.you[HON] AUX.PROG:1SG
chiedendo *solo* *la* *signora* ***lamenta*** *dolore* *da* *qualche*
ask:GER only ART.F lady experience:3SG pain at some
altra *parte*
other part
'Excuse me, excuse me, I am only asking if the lady experiences pain somewhere else.' (KIParla, KPS001)

b. *mio* *fratello* ***si*** ***lamenta*** *sempre*
my brother REFL complain:3SG always
'My brother is always complaining.' (KIParla, PBB024)

c. *lascio* *perdere* *quella* *st~* *storia* *perché* *lei* ***si*** ***lamenta*** ***di***
let:1SG lose:INF that st~ story because she REFL complain of
questa ***storia***
this story
'I leave that stuff alone because she complains about this stuff' (KIParla, KPS003)

Transitive verbs other than those illustrated in (7a–c) only permit deletion of the object, as illustrated in (7d–e):

(7) Italian

d. *caro* *rossi,* ***beve*** *un* *po'* *di* *grappa?*
dear R. drink:3SG[HON] a bit of grappa
'Dear Rossi, do you drink some grappa?' (KIParla, PTB026)

e. *c'* *è* *gente* *che* ***beve*** *per* *strada* *magari*
there be.3SG people REL drink:3SG through road possibly
che *fa* *un* *po'* *di* *rumore* *diciamo*
REL make.3SG a bit of noise so.to.speak
'There are people drinking in the street, maybe making a bit of noise let's say.' (KIParla, PBA020)

In the pattern illustrated in (7d–e) with *bere* 'to drink', the intransitive form in (7e) constitutes a case of agent-preserving lability, or ambitransitivity, in Dixon's (1994:

54 and passim) terms. Under this type of lability, both the transitive and intransitive uses of the verb retain an agentive argument. By contrast, in patient-preserving lability, both uses of the labile verb involve a patientive argument, as in *I broke the stick* vs. *The stick broke*. When the object is deleted, some transitive verbs may also develop a more specific, lexicalized interpretation (e.g. *bere* 'to drink alcohol'), a point to which we return below.

The phenomenon of object deletion in Italian has been investigated mostly within the generative tradition, where it has been treated as a structural issue rather than a purely pragmatic one. A central reference is Rizzi (1986), who argues that Italian allows syntactically represented null objects (pro) with arbitrary interpretation. His key argument is that deleted objects in Italian are syntactically active: for instance, they can control PRO and bind anaphors, a possibility that is systematically excluded in English. This contrast is explained through a parametric difference in the licensing of pro in object position. An important exception to this predominantly generative line of analysis is Cennamo (2017), who approaches object deletion from a semantic and pragmatic perspective. In particular, Cennamo (2017) argues that object deletion is not primarily syntactic, but depends on how arguments are licensed in the verb's event structure template and on the semantic content of the verb root. Objects licensed only as root participants, i.e. lexically entailed by the verbal root itself, are more easily deleted than event structure participants: for example, activity and consumption verbs such as *mangiare* 'eat' or *scrivere* 'write' allow object deletion, whereas achievements like *rompere* 'break' do not, unless used metaphorically (e.g. when *rompere* means 'be annoying'). Object deletion is also sensitive to aspectual boundedness (e.g. imperfective contexts favor deletion), animacy and definiteness of the object, and the agentivity of the subject: *uccidere* 'kill' allows object deletion more easily than the verb *assassinare* 'assassinate', whose subject is more agentive. Cennamo furthermore adopts the distinction between Indefinite and Definite Null Instantiation (Fillmore 1986), showing that the two phenomena are subject to different constraints. Indefinite Null Instantiation displays relatively stable cross-linguistic patterns and typically involves non-referential or weakly specified objects, as with activity and consumption verbs such as *mangiare* 'eat' or *bere* 'drink'. Definite Null Instantiation, by contrast, involves referential objects whose identity must be recoverable from the linguistic or situational context, as in anaphoric sequences (*Ho ascoltato la proposta e poi ho rifiutato* 'I listened to the proposal and then I refused (it)') or in imperatives (*Prendi!*, 'take (it)!'). This latter type in Italian is strongly

conditioned by discourse-pragmatic factors such as topicality and accessibility rather than by verb semantics alone.

The semantic and pragmatic parameters identified by Cennamo (2017) as favoring object deletion in Italian largely overlap with those considered in the present study. In particular, *object predictability* corresponds to the notion of *root participant*, i.e. an argument licensed by the lexical semantics of the verb rather than by its event-structural template. *Contextual accessibility* of the object provides a systematic account of object *recoverability from the linguistic or situational context*, as required in cases of Cennamo's (2017) Definite and Indefinite Null Instantiation. Moreover, *inherent atelicity* is one of the parameters already highlighted by Cennamo (2017) as facilitating object deletion, insofar as it shifts the focus from the affected participant to the event itself. To these factors, the present analysis adds the parameters of *verbal finiteness* and *agent affectedness*, which have not been systematically addressed in previous accounts. The innovative contribution of this study lies in the fact that these parameters are weighed according to their relative importance on the basis of a corpus of naturally occurring data, which has been annotated and quantitatively analyzed. Indeed, only a quantitative approach can provide a reliable basis for assessing the relative contribution of the different factors affecting object deletion in Italian.

In light of this background, the article aims to address the following research questions:

(i) What factors govern the alternation between the two possibilities exemplified in (7d-e) in spoken Italian?

(ii) Do the factors identified as relevant for APs across languages also influence object deletion in ambitransitive pairs in languages without an AP construction?

(iii) Are these factors of equal importance, or can they be ranked in terms of their influence?

By analyzing how these factors interact in naturally occurring spoken data, this study will enhance our understanding of the relationship between APs and object deletion constructions, and of how the latter can offer insights into the mechanisms that shape the use of APs in languages where such constructions exist. The findings will show that typological tendencies are deeply rooted in language use, and much of the cross-linguistic diversity can be meaningfully explained by combining large-scale typological surveys with the analysis of naturally occurring data in languages where this is feasible.

The article is structured as follows. Section 2 provides a brief overview of the data and methods used in this study. Section 3 discusses the factors identified as relevant to the likelihood of AP derivation across languages and presents a statistical analysis of their relative impact on the possibility of object deletion in spoken Italian. Section 4 offers some concluding remarks and considers the broader implications of the study.

2. Data and methods

The analysis is based on data extracted from two modules of the KIParla corpus (Mauri et al. 2019, 2.328.209 tokens), a recently developed resource for the study of spoken Italian. The KIParla corpus has a modular and incremental structure, comprising several internally organized subcorpora.[2]

For this study, the first two modules of the corpus (i.e., KIP and ParlaTO) were analyzed. These modules include data collected across different interactional contexts, involving a diverse range of speakers. More specifically, the KIP module contains recordings made in Bologna and Turin within university settings, including various types of interactions such as professor/student exchanges (lectures, office hours, and exams) and student/student interactions (semi-structured interviews and informal conversations). The ParlaTO module, by contrast, contains semi-structured interviews conducted in Turin with speakers representing a wide range of social backgrounds (e.g., age group, education level, profession). These interactional contexts can be classified according to the degree of power asymmetry between speakers. From this perspective, lectures, exams, and office hours involve clear asymmetry—characterized by more formal exchanges—whereas semi-structured interviews and free conversations tend to reflect a higher degree of symmetry, resulting in more informal interactions.

In order to create a dataset for the present study, a list of all transitive verbs attested in the corpus was extracted, excluding modal and auxiliary verbs. Once the frequency list was obtained, 4 frequency classes were identified, and 4 verbs were selected for each of them (see Table 1).

All occurrences of each verb were extracted and randomized. Then, the occurrences were manually cleaned by eliminating:

- Infinitive forms occurring as subject or direct object;

[2] The resource is freely accessible via the NoSketch Engine platform (www.kiparla.it).

- Passive and reflexive forms;
- Cases in which the direct object coincided with a complement clause;
- Cases in which a form of the verb was employed as a discourse marker (e.g. *guarda* 'look!' or *senti* 'listen!').

Frequency (f) class (*based on n. of occurrences*)	**Verb**	**Number of occurrences**
A	*capire* 'to understand'	1.471
f ≥ 1.000	*prendere* 'to take'	1.383
	trovare 'to find'	1.281
	guardare 'to watch'	1.051
B	*sentire* 'to hear'	976
1.000 > f ≥ 700	*scrivere* 'to write'	923
	chiamare 'to call'	906
	conoscere 'to know'	716
C	*cercare* 'to search for'	584
700 > f ≥ 400	*leggere* 'to read'	485
	finire 'to finish'	481
	studiare 'to study'	471
D	*mangiare* 'to eat'	384
400 > f ≥100	*aprire* 'to open'	275
	creare 'to create'	204
	bere 'to drink'	113

Table 1: Frequency list.

This preliminary work on the data led us to exclude occurrences of the verb *finire* 'to finish' because they were strongly context-dependent: in most cases, the verb occurred in interviews in which the interviewer reassured the interviewee with a fixed construction that the interview was about to end, as in (8).

(8) KIParla (PTA006)

TOR001: va bene, segui il calcio?
alright follow:2SG the football
'Okay, are you interested in football?'
TOI006: sì
yes
'Yes'

TOR001: *abbiamo* *quasi* *finito*
AUX:1PL almost finish:PST.PTCP
'We are almost done'

TOI006: *lo* *seguo […]* *molto* *attivamente* *il* *calcio*
OBJ follow:1SG very actively the football
'I am very interested in football'

The final dataset consists of the first hundred clean occurrences of each verb, for a total of 15 headwords and 1.500 occurrences. Then, each occurrence was manually annotated according to a series of linguistic predictors.

The dependent variable in the analysis is object realization, which has three possible values: zero, pronoun and noun phrase. Table 2 reports the frequencies.

Object realization	Absolute frequencies	Normalized frequencies
zero	331	22,1%
pronoun	308	20,5%
noun phrase	861	57,4%
Tot.	1.500	100%

Table 2: Object realization - frequencies.

Table 3 reports the distribution of values related to the realization (or non-realization) of the object and the verb lemmas in the dataset (listed in alphabetical order). Only absolute values are reported, as 100 occurrences were considered for each lemma.

Verbs	Object realization			
	zero	pronoun	NP	Tot.
aprire 'to open'	21	15	64	100
bere 'to drink'	48	6	46	100
capire 'to understand'	44	17	39	100
cercare 'to search for'	17	16	67	100
chiamare 'to call'	7	53	40	100
conoscere 'to know'	3	31	66	100
creare 'to create'	4	4	92	100
guardare 'to watch'	17	27	56	100
leggere 'to read'	25	27	48	100
mangiare 'to eat'	42	18	40	100
prendere 'to take'	3	25	72	100

Verbs	**Object realization**			
	zero	**pronoun**	**NP**	**Tot.**
scrivere 'to write'	25	14	61	100
sentire 'to hear'	10	25	65	100
studiare 'to study'	61	6	33	100
trovare 'to find'	4	24	72	100
Tot.	331	308	861	1500

Table 3: Verb lemmas - frequencies.

Then, taking into account factors that in the literature are considered to be relevant to the realization of the object (see Section 1), we considered semantic, syntactic and discourse-pragmatic predictors.

The semantic predictors are lexeme dependent:

- Inherent atelicity of the verb: this predictor has two values, yes (e.g. *bere*, 'to drink'), and no (e.g. *sentire*, 'to hear'). The frequencies are reported in Table 4.

Inherent atelicity	**Absolute frequencies**	**Normalized frequencies**
yes	800	53,3%
no	700	46,7%
Tot.	1.500	100%

Table 4: Inherent atelicity - frequencies.

- Agent affectedness: depending on the way in which the A-argument may be affected, we distinguish between physical affectedness (e.g. *mangiare*, 'to eat'), cognitive affectedness (e.g. *capire,* 'to understand), and no affectedness at all (e.g. *scrivere*, 'to write'). The observed frequencies are reported in Table 5.

Agent affectedness	**Absolute frequencies**	**Normalized frequencies**
physical	200	13,3%
cognitive	400	26,7%
no	900	60,0%
Tot.	1.500	100%

Table 5: Agent affectedness - frequencies.

- Object predictability: depending on the specific semantic feature of the object that may be predictable on the basis of the verb semantics, we distinguish

between predictable animacy (e.g. *aprire,* 'to open' > inanimate objects), predictable semantic field (e.g. *mangiare,* 'to eat' > food), no predictability (e.g. *guardare*, 'to watch' > anything); frequencies are reported in Table 6.

Object predictability	Absolute frequencies	Normalized frequencies
animacy	200	13,3%
semantic field	600	40,0%
no	700	46,7%
Tot.	1.500	100%

Table 6: Object predictability - frequencies.

Table 7 shows the annotation of the lexemes included in our dataset, ordered on the basis of their absolute frequencies in the corpus, for the three semantic predictors of inherent telicity, agent affectedness and object predictability:

Verb	Inherent atelicity	Agent affectedness	Object predictability
capire 'to understand'	no	cognitive	no
prendere 'to take'	no	no	no
trovare 'to find'	no	no	no
guardare 'to watch'	yes	no	no
sentire 'to hear'	no	no	semantic field (sounds)
scrivere 'to write'	yes	no	semantic field (texts)
chiamare 'to call'	no	no	no
conoscere 'to know'	yes	cognitive	no
cercare 'to search for'	yes	no	no
leggere 'to read'	yes	cognitive	semantic field (texts)
studiare 'to study'	yes	cognitive	semantic field (subjects)
mangiare 'to eat'	yes	physical	semantic field (food)
aprire 'to open'	no	no	animacy (inanimate)
creare 'to create'	no	no	animacy (inanimate)
bere 'to drink'	yes	physical	semantic field (liquids)

Table 6. Annotation of the verbs in the dataset, based on the three semantic predictors of Inherent atelicity, Agent affectedness and Object predictability.

The syntactic predictor concerns the finiteness of the verb form. It may have two possible values: finite (conditional, imperative, indicative and subjunctive), non-finite

(gerundive and infinitive). The frequencies observed for this parameter are reported in Table 8.

Finiteness	Absolute frequencies	Normalized frequencies
finite	1.266	84,4%
non-finite	234	15,6%
Tot.	1.500	100%

Table 8: Finiteness - frequencies.

At the discourse level, we considered object contextual accessibility: taking into account the immediately preceding context (≅ 100 words before), we distinguish between mention (the object has been explicitly mentioned in the preceding discourse, as in (9)), inference (the object is not mentioned, but is inferable from the preceding co-text, as in (10)), not-accessible (the object is not accessible in the preceding context, cf. Chafe 1994). Frequencies are reported in Table 9.

Object accessibility	Absolute frequencies	Normalized frequencies
mention	644	42,9%
inference	215	14,3%
not-accessible	641	42,7%
Tot.	1.500	100%

Table 9: Object contextual accessibility - frequencies.

In (9), the object of drink (i.e. *una birra*, 'a beer') is explicitly mentioned by TOI67, whereas in (10) the object of drink (*acqua,* 'water') is never mentioned but is inferable from the frame of swimming, activated in the previous turns.

(9) KIParla (PTA005)

TOR001: *cioè volete **una birra?***
I.mean want:2PL *a beer*
'I mean, do you want a beer?'
TOI005: *no, se la vuoi aprire **la bevo** poi*
no if 3SG.OBJ want:2SG open:INF 3SG.OBJ drink:1SG later
'No, if you want to open it, I will drink it later'

TOR001: *a questo punto*
at this point
'At this stage'
TOI005: *già che ci siamo*
already COMP LOC be:1PL
'Since we are here'
TOR001: xxx *scusa?*
sorry
'xxx sorry?'
TOR002: *eh*?
??
'What?'
TOR001: *dico, siam qua* ***beviamoci*** ***una birra***
I.mean be:1PL here drink:1PL:REFL a beer
'I mean, since we are here, let us have a beer'

(10) KIParla (TOD2010)

TO046 *non ho mai imparato a nuotare in stile libero*
not have:1SG ever learn:PST.PTCP to swim:INF in style free
'I have never learned to swim freestyle'
TO055 *ma come non hai mai impara~ è il primo*
but how not have.2SG ever learn:PST.PTCP be.3SG the first
stile che si impara
style REL REFL learn:3SG
'But how come you've never lear— it's the first style one learns'
TO046: *non ho mai imparato*
not have:1SG ever learn:PST.PTCP
'I've never learned'
TO046: *ma avevo fatto i corsi da bambino come tutti*
but have:IPFV.1SG do:PST.PTCP the courses as child like all
'But I had taken swimming classes as a child, like everyone'
TO046: *però non ho mai imparato*
however not have:1SG ever learn:PST.PTCP
'however I never learned'

TO046: *ho dei problemi a respirare perché faccio due*
have:1SG some problems to breathe:INF because do:1SG two
bracciate e poi emergo e poi emergo male
strokes and then surface:1SG and then surface:1SG badly
*e quindi **bevo** **acqua***
and so drink:1SG water
'I have trouble breathing because I do two strokes and then I surface, and then I surface badly, and so I swallow water'

After the annotation, in order to verify the impact of the different linguistic factors on object realization, a statistical analysis was conducted using a conditional inference tree and random forest (Levshina 2015). A conditional inference tree is a decision tree that selects splitting variables, which partition data by recursively applying hypothesis tests, ensuring statistically sound splits. This results in a model that is both interpretable and robust against biased variable selection. A random forest consists of an ensemble of conditional inference trees to improve prediction accuracy and model stability. Each tree in the forest is trained on a random subset of the data, and random feature selection is applied at each split. By aggregating diverse trees, random forests reduce variance and are less prone to overfitting compared to single trees. They are recommended in cases in which there are few linguistic occurrences and many predictors, and they can work with unbalanced data (for a thorough discussion see Tagliamonte & Baayen 2012 and Levshina 2015: 291-300). Furthermore, conditional inference trees and random forests can effectively handle factors interactions without requiring you to specify them explicitly, as they naturally capture interactions between variables by splitting the data at different levels.

Finally, we opted to complement these approaches with a mixed-effects logistic regression, since inference tree and random forest offer limited inferential transparency with respect to the direction, magnitude and statistical significance of individual predictors. A mixed-effects logistic regression allows for a more explicit hypothesis-driven assessment of fixed effects, while simultaneously accounting for structured random variability in the data. In particular, the inclusion of random effects makes it possible to control for lexical dependencies. This modelling strategy thus combines the exploratory strengths of machine-learning approaches with the inferential rigor required for theory-driven analysis.

The whole analysis was conducted using R and made use of the following packages: readxl, party, Hmisc and lme4.

3. Results: assessing the factors at play

This section is devoted to the analysis of the data.

As a first step, we examine the joint distribution of two parameters, namely the realization and the accessibility of the object (see Figure 1), since we hypothesize a strong relationship between these two dimensions.

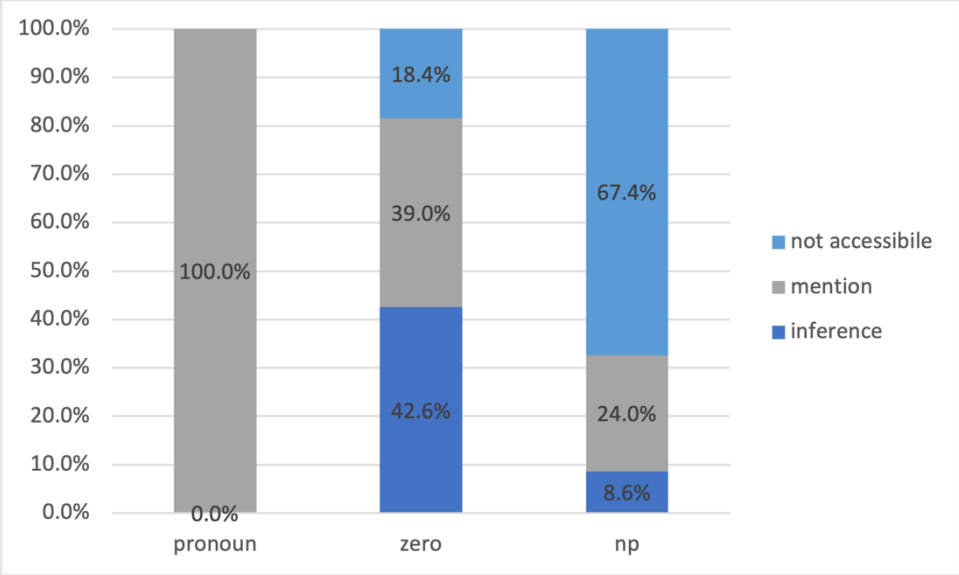


Figure 1: Object-realization and accessibility

As shown in the figure, object deletion (*zero*) is most frequent when the object has been mentioned in the preceding co-text (39%) or is inferable (42.6%). By contrast, when the object is not accessible, it is predominantly realized as a full NP (67.4%). Both *zero* and NP realizations are not restricted to a single accessibility condition: each of them is attested across all values of contextual accessibility, albeit with different distributions. This contrasts with pronouns, which show no such variability: when the object is realized pronominally, it is categorically mentioned in the preceding co-text.

This pattern is illustrated in example (11). *Una tua attività* 'your own business' functions as the object of both *hai* 'you have' and *aprire* 'to open': it is first overtly introduced and subsequently referred to by the clitic pronoun *la*, attached to *aprire*.

(11) KIParla (PTA001)

TOI001: *e poi eh tutto dipende uno cosa vuole fare nella vita*
and then eh all depend:3SG one what want:3SG do:INF in:DEF life
'and then ehm it all depends on what one wants to do in life'
eh se hai una tua attività ***aprir-la*** *qua o*
eh if have:2SG INDEF your activity open-3SG.OBJ here or
aprir-la all'estero
open-3SG.OBJ abroad
'eh, if you have your own business, to **open it** here or open it abroad…'

For this reason, occurrences in which P is realized through a pronoun were excluded from the quantitative analysis. The dataset considered therefore consists of 1.192 occurrences.

3.1. Conditional inference tree

Figure 2 shows the conditional inference tree of object deletion, resulted by the statistical analysis conducted on our dataset. The c-index of the model is 0,87 and thus has excellent discrimination (Hosmer & Lemeshow 2000: 162).

The first parameter that proves to be significant in splitting the data into internally homogeneous subsets is object contextual accessibility (node 1). When the object is not accessible in the preceding context, the model identifies affectedness as the next most significant factor (node 2). In particular, when the agent is not affected, the object is almost invariably realized overtly (node 5). Zero realization is instead possible for non-accessible objects with verbs involving some degree of agent affectedness. If the agent is affected at the physical level (as with *mangiare* 'eat' and *bere* 'drink'), the object shows a greater tendency to be deleted even in the absence of prior contextual accessibility (node 3, ca. 40% of the occurrences).

Example (12) provides a case in point: the object of *abbiamo sempre mangiato* 'we have always eaten' is not accessible in the preceding context, but it is any case deleted. The verb 'to eat' is indeed relevant in the discourse context for how it affected the agent with positive consequences (i.e. nourishment and survival), rather than for the specific properties of the patient.

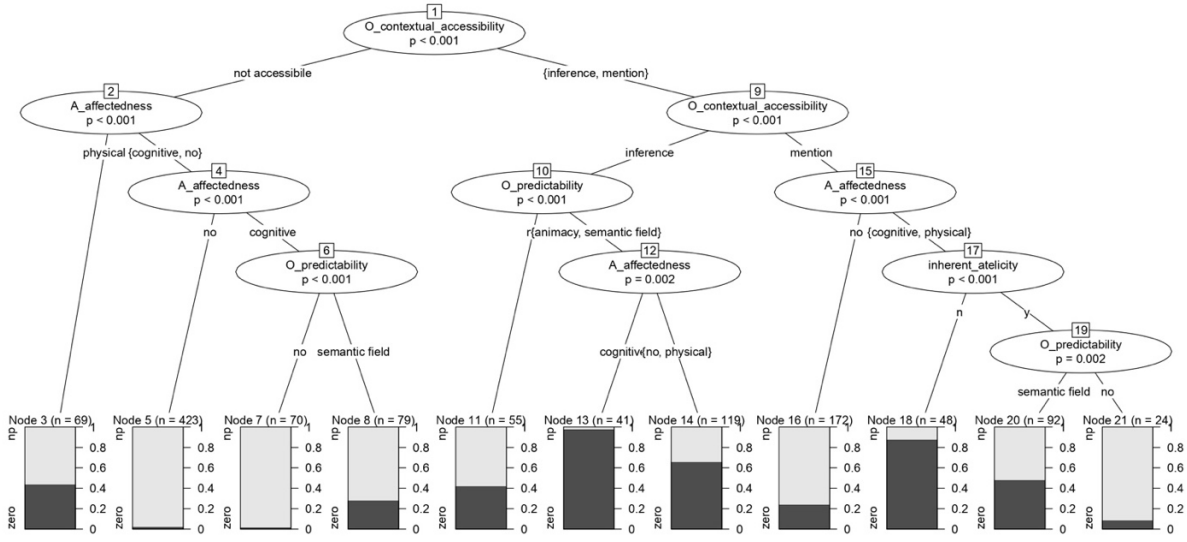


Figure 2: Conditional Inference tree.

Actually, the identification of the specific foods that were eaten is irrelevant and even problematic, given the plurality of specific events of eating being referred to (cf. the use of the adverb *sempre* ‘always’).

(12) KIParla (PTB018)

TOI056:[…] *quindi io n~ non mi vergogno di dire*
so I NEG REFL be.ashamed:1SG of say:INF
‘[…] So I am not ashamed in saying (that) ’
no~ ***abbiamo sempre mangiato*** *diciamo*
1PL AUX:1PL always eat:PST.PTCP let’s.say
‘We have always eaten, let’s say’
eh ci è sempre stato il
eh LOC AUX.3SG always be:PST.PTCP DEF
‘There has always been the’
l' indispensabile in casa però periodi difficili ce ne
DEF indispensable at home but moment:PL difficult:PL LOC PART
sono stati
AUX.3PL be:PTCP.PST
‘The indispensable at home, but difficult moments we had a few’

The tendency to delete non-accessible objects is reduced when the agent is only cognitively affected. Within this latter group, object predictability (node 6) further refines the distribution, showing that zero realization of non-accessible objects is admitted when they are highly predictable, as they belong to a semantic field required by the verb's lexical semantics (node 8).

By contrast, when the object is accessible in the preceding context, the tree reveals a more articulated structure. If accessibility is achieved through inference alone, object predictability comes into play (node 10): when no semantic feature of the object can be predicted from the verb's lexical semantics, the object is more frequently realized as a full noun phrase; conversely, when the verb constrains the object's animacy or semantic field, deletion becomes more likely, even though affectedness still plays a role (node 12).

When the object is accessible because it has been overtly mentioned in the preceding context, affectedness again emerges as a significant discriminator (node 15), while inherent atelicity (node 17) and object predictability (node 19) further modulate realization preferences.

Overall, the inference tree shows that object realization is governed by a complex interplay between contextual accessibility and verb semantics, especially with respect to agent affectedness and the semantic constraints that make the object's animacy or semantic field predictable. Contextual accessibility constitutes the primary organizing principle: objects that are not accessible in the preceding context tend to be realized by means of full NPs, while deletion becomes increasingly available as accessibility and semantic predictability increase. Examples (13) and (14) may help us understand the mechanisms at work.

(13) KIParla (BOD2014)

BO115:	*cioè*	*se ha*	*bisogno*	*di parlare*	*robe*	*così*	
	I.mean	if have:3SG	need	of talk	things	so	
	'I mean, if he needs to talk or stuff like that'						
	glie-lo	*dico*	*sempre*				
	3SG.DAT-3M.SG.OBJ	tell:1SG	always				
	'I always tell him'						
	scrivimi	*se hai*	*bisogno*				
	write:1.SG.DAT	if have:2SG	need				
	'«Write to me if you need»'						

però	*non*	*l'*	*ha*	*mai*	*fatto*	*capito*
but	NEG	3SG.OBJ	AUX:3SG	never	do:PST.PTCP	you.know

'But he never did it, you know'

In this free conversation between friends, the object of *scrivere* is easily inferable from context: it can be a message, a text, or a similar informal way of contacting a person on the phone to fix an appointment or to talk about problems. In this case, there is clear focus on the relation between the agent and the recipient, while the specific P-argument is not relevant. In example (14), on the other hand, the context is office hours, and the student is discussing some issues with the professor. In this case too, the relevant arguments are the beneficiary and the agent, while the patient is easily accessible from the context and does not need to be syntactically encoded. The object of *scrivere* in this example cannot be an informal message on the phone, but it is necessarily an email, namely, the only object that could be written to the Student Office.

(14) KIParla (BOA1018)

BO082:	*okay*	*ma*	*l'*	*ufficio*	*didattico*	*o*	*la*	*segreteria*	*studenti?*
	okay	but	DEF	office	teaching	or	DEF	secretariat	student.PL

'Okay, but the Teaching Office or the Students Office?

BO085:	*okay*	*io*	***ho***	***scritto***	*a*	*entrambi*
	okay	I	AUX:1SG	write:PST.PTCP	to	both

'Okay, I wrote to both'

The object is thus frequently inferred from the specific speech situation. For instance, when two university students are engaged in a casual conversation, their use of *cercare* 'to look for' is likely to refer to searching for housing, *trovare* 'to find' typically relates to job hunting, and *scrivere* 'to write' most often refers to sending short messages on the phone, as in (13). As a result, explicitly specifying the object of verbs like 'look for', 'find', and 'write' can become unnecessarily redundant, and the object is simply deleted.

In some cases, accessibility arises when the discourse context activates a culture-specific frame, such as nightlife or moving to a new city: the object is perceived by

interlocutors as obvious, leading to a shift in focus from the object to the activity itself. We also observe instances of semantic specialization (cf. Wilson 2003) resulting from repeated use within specific frames. For example, in the frame of a night out with friends (as in *movida* in example (15)), *bere* ∅ 'drink ∅' comes to denote the act of 'drinking alcohol', whereas in the context of a medical check-up, it may refer instead to water and proper hydration.

(15) KIParla (PTA003)

TOI003: *ma parli di movida?*
but talk:2SG of movida
'Are you talking about nightlife?'
TOR001: *sì, di Torino*
yes of Turin
'Yes, of Turin'
TOI003: *di movida?*
of nightlife
'Nightlife?'
TOR001: *mh mh*
mh mh
'Mh mh'
TOI003: *allora sicuramente piazza vittorio per* ***bere***
then definitely Piazza Vittorio to drink:INF
'Then definitely Piazza Vittorio for **a drink**'

Interestingly, some of the examples of APs discussed in the literature point to similar phenomena. Example (1b) from Chukchi (cf. Section 1) shows that the AP form of 'take' is narrowed to the specialized reading 'taking/winning a prize', based on a culture-specific frame that makes the object realization unnecessary. Example (4b) from Tolowa (see Section 1.1) provides an instance of the AP form of 'drink' that specializes as 'drink alcohol', likely as a consequence of a repeated frame in which the object of drinking is obvious and thus easily deleted, as in (15).

3.2. Random forest

Figure 3 shows the predictors considered in the analysis and their relative importance in influencing object realization (C index = 0.88, excellent discrimination, see Hosmer & Lemeshow 2000: 162).

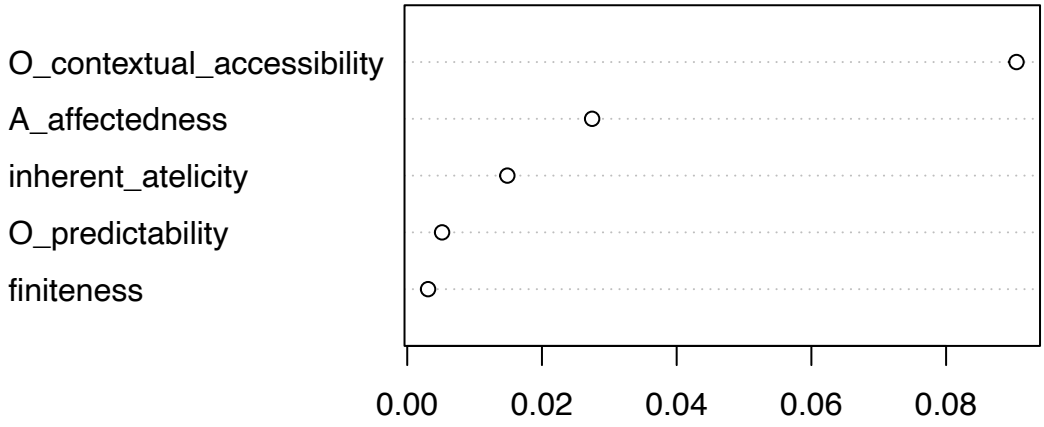


Figure 3: Random forest.

A comparison with Figure 1 reveals a high degree of correspondence: : predictors that emerged as relevant in the inference tree also rank high in the random forest.

The most important predictor for object realization is object CONTEXTUAL ACCESSIBILITY. Its distribution is reported in Table 10, which provides both absolute frequencies and percentages. The association between contextual accessibility and object realization is statistically significant ($\chi^2 = 278{,}6375$, p value $< 0{,}00001$).

CONTEXTUAL ACCESSIBILITY	**Zero**	**NP**
inference	141 (42,6%)	74 (8,6%)
mention	129 (39,0%)	207 (24,0%)
not-accessible	61 (18,4%)	580 (67,4%)
tot	1.171 (100%)	329 (100%)

Table 10. Contextual accessibility and object realization.

The distribution shows that the object is most frequently realized as a full NP when it is not accessible in the preceding discourse (67,4%), less often when it has been overtly mentioned in previous context (24,0%) or is accessible by inference alone (8,6%). A complementary pattern is observed for zero realization, which is predominantly attested when the object is accessible in the preceding context, either

through inference (42,6%) or via prior mention (39,0%), and only marginally when the object is not accessible (18,4%).

The second most important parameter is AGENT AFFECTEDNESS, whose distribution is reported in Table 11. This distribution is likewise statistically significant (χ^2 = 131,0982, p value < 0,00001).

AGENT AFFECTEDNESS	**Zero**	**NP**
physical	86 (27,2%)	90 (10,0%)
cognitive	186 (40,2%)	133 (21,6%)
no	589 (32,6%)	108 (68,4%)
tot	861 (100%)	331 (100%)

Table 11. Agent affectedness and object realization.

The data reveal systematic differences between the two realization types. Zero objects are more frequently associated with affected agents, particularly in the cognitive category (40,2%), while approximately one third of the occurrences involves verbs with no agent affectedness (32,6%). Objects realized as full noun phrases, by contrast, occur predominantly with verbs involving no agent affectedness (68,4%), and only marginally with cognitively (21,6%) or physically (10%) affected agents. These patterns suggest an association between higher degrees of agent affectedness and zero object realization, while overt realization by a full NP is favored when the agent is not affected.

Corpus data thus support the view that object deletion is often a consequence of the reduced relevance of the P-argument with respect to the A-argument and to the activity itself: when the agent undergoes a change in state and coincides with the activity endpoint, explicit identification of the patient becomes less necessary in discourse.

The third predictor in the random forest ranking is INHERENT_ATELICITY (Fisher exact test statistic value is < 0,0001, the result is significant at p >0,01): object deletion is favored in actions that do not imply a specific endpoint, but whose telicity is determined by the incremental nature of the object (see Section 1.1). In such cases, the discourse focus readily shifts from the patient to the action itself. As shown in Table 12, 71,9% of zero-object occurrences involve atelic predicates, while realized-object cases are more evenly distributed across telic and atelic verbs.

INHERENT ATELICITY	**Zero**	**NP**
yes	238 (71,9%)	417 (48,4%)
no	93 (28,1%)	444 (51,6%)
tot	331 (100%)	861 (100%)

Table 12. Inherent atelicity and object realization.

3.3. Logistic regression

The logistic regression yielded similar results, while also highlighting some additional aspects. A generalized linear mixed-effects model with a binomial link function was fitted to predict OBJECT REALIZATION. To account for lexical variability, random intercepts for the verb lexeme were included. Unfortunately, we could not include speakers as a random effect; future studies may address this aspect. The model was estimated using maximum likelihood with Laplace approximation.

The model yields an AIC of 1070,4 and a log-likelihood of -525,2, based on 1.283 observations and 15 lexical types. The random-effects structure reveals substantial variability across verb lexemes: the random intercept for LEXEME shows a variance of 0,623 (SD = 0,790), indicating meaningful differences in baseline probabilities of object realization across verbs. Figure 4 illustrates the distribution of NP and zero object realizations across verb lemmas (see Table 3 for the complete picture).

Turning to the fixed effects, the intercept is negative and statistically significant (Estimate = −1,692, SE = 0,773, z = −2,189, p = 0,0286), corresponding to a low baseline probability of object realization for the reference levels of all predictors.

INHERENT ATELICITY does not exert a reliable effect once other predictors are controlled for (Estimate = −0,109, SE = 0,572, z = −0,190, p = 0,849). Similarly, predictors related to OBJECT PREDICTABILITY do not reach statistical significance, both for lack of predictability (Estimate = 0,960, SE = 0,662, z = 1.450, p = 0.147) and for semantic-field predictability (Estimate = 0,099, SE = 0.806, z = 0.123, p = 0.902). FINITENESS also shows no significant effect: non-finite clauses show a positive but non-significant tendency toward higher object realization (Estimate = 0,286, SE = 0,218, z = 1,312, p = 0,189).

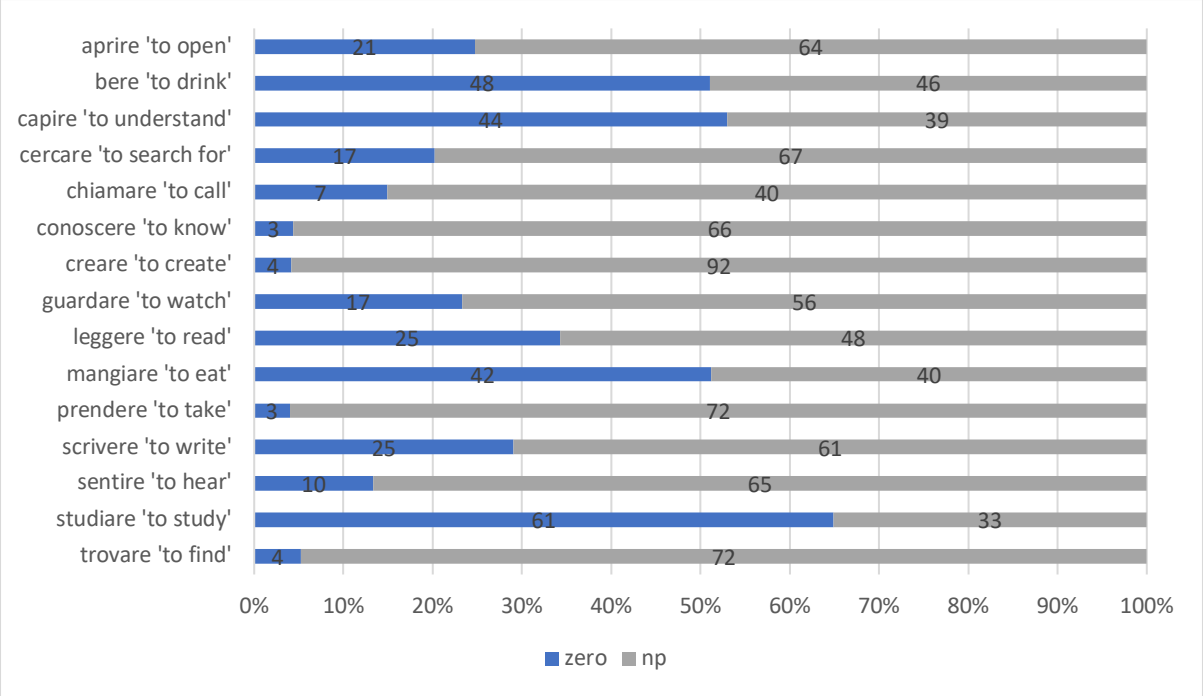


Figure 4. Lexical variability in object realization

By contrast, AGENT AFFECTEDNESS shows a significant effect: the absence of affectedness significantly increases the likelihood of overt object realization relative to the reference level (Estimate = 1m213, SE = 0,569, z = 2,131, p = 0,033), corresponding to a substantial increase in the odds of realization. Physical affectedness does not differ significantly from the baseline (Estimate = 0.059, SE = 0.777, $z = 0.076$, $p = 0.940$), indicating no systematic effect. OBJECT CONTEXTUAL ACCESSIBILITY emerges as the strongest predictor in the model. Objects that are accessible (by mention or inference) are significantly less likely to be realized overtly (Estimate = 1,238, SE = 0,202, z = 6.127, $p < 0.001$). An even larger effect is observed for non-accessible objects, which show a dramatic increase in the log-odds of realization (Estimate = 3,089, SE = 0,214, z = 14,428, p < 0.001). Crucially, these effects remain robust even after accounting for random variation across lexemes.

Overall, the mixed-effects analysis confirms the central role of discourse-level factors, most notably object contextual accessibility, in predicting object realization. At the same time, it shows that aspectual properties, object predictability, and finiteness do not contribute significantly once lexical differences and accessibility are taken into account. The inclusion of random intercepts for verb lexeme further demonstrates that object realization is sensitive to both discourse constraints and item-specific baseline tendencies.

4. Conclusions and future prospects

This study has investigated object deletion in spoken Italian as a discourse strategy for backgrounding the patient argument, and has assessed to what extent the factors that typological studies have identified for AP constructions also shape object deletion in a language without dedicated AP morphology. Using naturally occurring corpus data and a combination of interpretable and predictive modelling techniques (conditional inference tree, random forest, and mixed-effects logistic regression), we have shown that object realization in Italian is governed by a complex interplay of discourse accessibility and verb-related constraints, alongside substantial lexeme-specific variability.

A first robust result concerns the primacy of object contextual accessibility. Across modelling approaches, accessibility emerges as the strongest predictor of whether the object is overtly realized or deleted. When the object is not accessible in the preceding discourse, overt realization as a full NP becomes the default option; conversely, deletion is favored when the object is already available in the common ground, either because it has been mentioned or because it can be inferred from the ongoing discourse. This finding matters for typology because contextual accessibility is rarely operationalized in cross-linguistic work on APs, yet it provides a principled account of when demoting or omitting the patient becomes discourse-economical.

Second, agent affectedness plays a systematic role in shaping object realization preferences, although its contribution is best understood as modulatory rather than categorical. In the inference tree, affectedness refines the distribution especially when the object is not contextually accessible, making deletion more available with verbs that involve some degree of agent affectedness (notably consumption predicates), while strongly disfavoring deletion when the agent is not affected. The mixed-effects model corroborates this general asymmetry by showing that the absence of affectedness increases the odds of overt object realization. Taken together, these results support the idea that patient backgrounding is facilitated when the event structure foregrounds changes in, or consequences for, the agent rather than outcomes for the patient.

Third, while inherent atelicity and object predictability emerge as relevant dimensions in the tree/forest analyses, they do not yield reliable independent effects in the mixed-effects model once contextual accessibility and lexical variability are

taken into account. This suggests that their contribution is largely indirect: they may matter insofar as they co-occur with particular verb classes, interactional routines, and discourse settings that make objects easier to recover or easier to treat as backgrounded. Rather than treating these dimensions as uniform predictors of object deletion, the evidence points to a scenario in which they primarily structure *where* deletion becomes pragmatically licensed and recurrent.

Finally, the mixed-effects analysis highlights a key methodological and theoretical point: object realization is characterized by substantial lexeme-specific baseline differences, even after controlling for the predictors considered. This aligns well with typological observations that APs are frequently lexically restricted across languages. The Italian data thus support a broader view in which patient-backgrounding strategies are not only constructional options, but also reflect the propensity of individual predicates to support economical reference tracking in discourse.

Beyond these quantitative results, the qualitative examination of corpus data shows how recurrent discourse frames and speech situations can make the object effectively obvious, thereby encouraging deletion and, in some cases, facilitating semantic specialization (e.g., 'drink' in nightlife-related frames). Importantly, the cross-linguistic parallels often noted for AP-derived specializations suggest that repeated deletion of predictable objects can be a general pathway linking argument realization, discourse routines, and meaning change.

Overall, our findings provide converging evidence that object deletion in an accusative language like Italian can perform functions that are closely comparable to those of AP constructions cross-linguistically, even in the absence of dedicated morphology. At the same time, they refine the typological picture by showing that discourse accessibility is a primary organizing principle for patient backgrounding in spontaneous speech, while verb semantics and aspectual properties act as secondary constraints whose effects are partly mediated by lexical and contextual factors.

Future research should (i) broaden the lexical coverage to include a larger set of transitive predicates, allowing finer-grained generalizations about verb classes and the distribution of 'natural' patient-backgrounding verbs; (ii) extend the annotation to interactional and sociolinguistic variables (e.g. speech event type, degree of formality, participant roles), which are likely to affect shared knowledge and hence object recoverability; and (iii) pursue richer random-effects structures (including speakers where feasible) to better model discourse and individual variability. Ultimately, this research agenda contributes to a usage-based, discourse-sensitive

typology of argument realization, in which spoken data provide crucial evidence for understanding how patient-backgrounding strategies are conditioned, routinized, and potentially conventionalized.

Acknowledgements

This research was developed within the within the framework of "DiverSIta – Diversity in spoken Italian" project, funded by the Italian Ministry of University and Research under the PRIN 2022 PNRR Call (PI: Caterina Mauri; n. P2022RFR8T; CUP J53D23017320001). We would like to thank two anonymous reviewers for their insightful comments, which helped us make the argumentation and the methodology sounder.

Abbreviations

> = acts on
1 = 1st person
2 = 2nd person
3 = 3rd person
A = agent
ABS = absolutive
AP = antipassive
ART = article
AUX = auxiliary
COMP = complementizer
COP = copula
D = multifunctional *d-* prefix (in Athabaskan)
DAT = dative
DEF = definite
ERG = ergative
F = feminine
GER = gerund
HON = honorific form
INF = infinitive
IPFV = imperfective
LOC = locative
M = masculine
N = uninterpreted morpheme (in Kuku Yalanji)
NEG = negation
NPST = non-past
OBJ = object
OBL = oblique
PART = partitive clitic
PL = plural
POS = positive
PRS = present
PROG = progressive
PST = past
PT = 'potent' case inflection (in Kuku Yalanji)
PTCP = participle
RED = reduplication
REFL = reflexive
REL = relativizer
SBJ = subject
SBJV = subjunctive
SG = singular
SIMULT = simultaneous
TH = them

References

Blake, Barry J. 1982. The absolutive: Its scope in English and Kalkatungu. In Paul J. Hopper & Sandra A. Thompson (eds.), *Studies in transitivity*, 71–94. New York: Academic Press.

Cennamo, Michela. 2017. Object omission and the semantics of predicates in Italian in a comparative perspective. In Lars Hellan, Andrej Malchukov & Michela Cennamo (eds.), *Contrastive studies in verbal valency*, 251–273. Amsterdam: John Benjamins.

Chafe, Wallace. 1994. *Discourse, consciousness, and time. The flow and displacement of conscious experience in speaking and writing*. Chicago: The University of Chicago Press.

Coon, Jessica. 2013. *Aspects of split ergativity*. Oxford: Oxford University Press.

Cooreman, Ann. 1994. A functional typology of antipassives. In Barbara Fox & Paul J. Hopper (eds.), *Voice: Form and function*, 49-88. Amsterdam: John Benjamins.

Creissels, Denis. 2012. The origin of antipassive markers in West Mande languages. Paper presented at the 45th Annual meeting of the Societas Linguistica Europaea (SLE). Stockholm, Sweden.

Creissels, Denis. 2015. Valency properties of Mandinka verbs. In Andrej Malchukov & Bernard Comrie (eds.), *Valency classes in the world's languages*, Vol. 1, 221–260. Berlin: Mouton de Gruyter.

Dixon, R. M. W. 1994. *Ergativity*. Cambridge: Cambridge University Press.

Dixon, R. M. W. & Alexandra Aikhenvald. 2000. Introduction. In R. M. W. Dixon & Alexandra Aikhenvald (eds.), *Changing valency: Case studies in transitivity*, 1–29. Cambridge: Cambridge University Press.

Dunn, Michael John. 1999. *A grammar of Chukchi*. Canberra: Australian National University. (Doctoral Dissertation.)

Givón, Talmy & Loren Bommelyn. 2000. The evolution of de-transitive voice in Tolowa Athabaskan. *Studies in Language* 24 (1). 41–76.

Heaton, Raina. 2017. A typology of antipassives, with special reference to Mayan. Mānoa: University of Hawai'i at Mānoa. (Doctoral Dissertation.)

Hosmer, David W., Jr. & Stanley Lemeshow. 2000. *Applied logistic regression*. 2nd ed. New York: Wiley.

Janic, Katarzyna. 2021. Variation in the verbal marking of antipassive constructions. In Katarzyna Janic & Alena Witzlack-Makarevich (eds.), *Antipassive. Typology, diachrony, and related constructions*, 249–291. Amsterdam: John Benjamins.

Khalilova, Zaira & Bernard Comrie. 2013. Bezhta. In Iren Hartmann, Martin Haspelmath & Bradley Taylor (eds.), *Valency Patterns Leipzig*. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at http://valpal.info/contributions/bezh1248).

Levshina, N. 2015. *How to do linguistics with R*. Amsterdam: Benjamins.

Mauri, Caterina & Silvia Ballarè. In press. *To understand in interaction: the rise of epistemic and evidential constructions based on* capire *in spoken Italian*. In Karolina Grzech & Henrik Bergqvist (eds.), *Expanding the boundaries of epistemicity*. Berlin: Mouton de Gruyter.

Mauri, Caterina, Silvia Ballarè, Eugenio Goria, Massimo Cerruti & Francesco Suriano. 2019. KIParla corpus: a new resource for spoken Italian. In Raffaella Bernardi, Roberto Navigli & Giovanni Semeraro (eds.), *Proceedings of the 6th Italian Conference on Computational Linguistics CLiC-it*. (URL: http://ceur-ws.org/Vol-2481/paper45.pdf)

Patz, Elisabeth. 2002. *A grammar of the Kuku Yalanji language of North Queensland*. Canberra: Pacific Linguistics.

Polinsky, Maria. 2013. Antipassive constructions. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures online*. Leipzig: The Max Planck Institute for Evolutionary Anthropology. Available online at: https://wals.info/chapter/108.

Polinsky, Maria. 2017. Antipassive. In Jessica Coon, Diane Massam, & Lisa deMena Travis (eds.), *The Oxford handbook of ergativity*, 307–331. Oxford: Oxford University Press.

Rizzi, Luigi. 1986. Null objects in Italian and the theory of pro. *Linguistic Inquiry* 17 (3). 501-557.

Sansò, Andrea. 2017. Where do antipassive constructions come from? A study in diachronic typology. *Diachronica* 34 (2). 175–218.

Sansò, Andrea. 2018. Explaining the diversity of antipassives: Formal grammar vs. (diachronic) typology. *Language and Linguistics Compass* 12 (6). 1–22.

Sapién, Racquel-María, Natalia Cáceres Arandia, Spike Gildea & Sérgio Meira. 2021. Antipassive in the Cariban family. In Katarzyna Janic & Alena Witzlack-Makarevich (eds.), *Antipassive. Typology, diachrony, and related constructions*, 65–96. Amsterdam: John Benjamins.

Say, Sergey. 2021. Antipassive and the lexical meaning of verbs. In Katarzyna Janic & Alena Witzlack-Makarevich (eds.), *Antipassive. Typology, diachrony, and related constructions*, 177–212. Amsterdam: John Benjamins.

Tagliamonte, Sali A. & R. Harald Baayen. 2012. Models, forests, and trees of York English: *Was/were* variation as a case study for statistical practice. *Language Variation and Change* 24 (2). 135-178.

Wilson, Deirdre. 2003. Relevance and Lexical Pragmatics. *Italian Journal of Linguistics* 15.2. 273-291.

CONTACT
silvia.ballare@unibo.it
caterina.mauri@unibo.it
asanso@unisa.it

Linguistic Typology at the Crossroads 5-2 (2025): 323-351 ISSN 2785-0943

Choice and complexity: In naturally occurring data, absolute complexity does not necessarily trigger relative complexity

THOMAS VAN HOEY[1,2], BENEDIKT SZMRECSANYI[2], MATT H. GARDNER[3]
[1] FWO, [2] KU LEUVEN, [3] QUEEN MARY UNIVERSITY OF LONDON

Submitted: 05/06/2024 Revised version: 04/04/2025
Accepted: 10/04/2025 Published: 25/02/2026

Abstract

This article interrogates two related assumptions widespread in many approaches to language: (1) languages do not like synonymy; (2) absolute complexity (i.e. the length of the grammatical description of a language) tends to be proportional to relative complexity (i.e. difficulty). Against this backdrop, we explore the link between syntactic synonymy (i.e., grammatical variation and optionality) and relative complexity (i.e., cognitive load) using methods from both corpus and psycholinguistics. We test two predictions: First, if synonymy avoidance is a design feature of human language, then grammatical variation should be sub-optimal and cause a measurable increase in production difficulty. Second, optionality will necessarily increase the absolute complexity of a language system. This increased absolute complexity will, in turn, increase relative complexity, i.e., cognitive load, also measured by increased production difficulty. Contrary to these predictions, analyses based on the SWITCHBOARD corpus of American English shows that the presence of choice contexts does not positively correlate with two metrics of production difficulty, namely filled pauses (*um* and *uh*) and unfilled pauses (speech planning time), not even when a typology of grammatical alternation type (insertion/deletion, substitution, permutation) is taken into account. These results challenge the view that grammatical optionality is sub-optimal and difficult for speakers, and that absolute complexity is necessarily proportional to relative complexity.

Keywords: variation; dysfluency; isomorphism; alternation; repeated measures correlation; SWITCHBOARD.

1. Introduction

Using naturalistic, conversational corpus data and focusing on "structurally and/or lexically different ways to say functionally very similar things" (Gries 2017: 8; see also Labov 1972a: 188), this paper interrogates two widespread assumptions in many approaches to language, ranging from functional linguistics to corpus linguistics and beyond: (1) synonymy and optionality are sub-optimal and (2) absolute language complexity is a determinant of relative language complexity. More specifically:

(1) Languages and language users disfavor synonymy and form-function asymmetries, and favor isomorphism instead.
(2) Relative language complexity is proportional to absolute language complexity.

While these two core ideas are prevalent in the general thinking about language (see a recent reaction against this idea by Wälchli & Sjöberg 2025 in this journal), ranging from functional linguistics to corpus linguistics and beyond, they are usually seen as opposite to fundamental concepts in variationist (socio)linguistics. Variationist linguists approach language from the position that there are times in which an identical meaning or grammatical function can be expressed using different formal means, be it in terms of lexical choice (3), grammatical construction (4), or phonological form (5). According to variationist linguistics, when people speak, sign, or write, they will need to make choices between "alternate ways of saying the same thing" (Labov 1972a: 188).

(3) Using *awesome/great/brilliant* etc. for highly positive evaluations (Tagliamonte & Pabst 2020)
(4) Using a synthetic or analytic construction for comparative adjectives, e.g., *happier than* vs *more happy than* (Suikkanen 2018)
(5) Using a released or unreleased plosive word-finally: *east* [istʰ] vs [ist̚] (Eckert 2008)

The primary data source in variationist linguistics typically consists of natural or naturalistic corpus data (Labov 1972b; 1984; Szmrecsanyi 2017), although experimental approaches are also used (see, e.g., Bresnan & Ford 2010). Patterns of variation, like those in (3)-(5), often delineate groups of speakers (interspeaker

variation); however, optionality is also typical of the speech of a single individual (intraspeaker variation). In this paper, we make use of a corpus that samples naturalistic dialogic speech to investigate options that are, in principle, available to all speakers of American English (so-called type-3 variability by van Hout & Muysken 2016).

The variationist approach is at odds with constructionist postulates, such as Goldberg's (1995: 67) Principle of No Synonymy:

> *The Principle of No Synonymy*: If two constructions are syntactically distinct, they must be semantically or pragmatically distinct (cf. Bolinger 1968; Haiman 1985; Clark 1987; MacWhinney 1989). Pragmatic aspects of constructions involve particulars of information structure, including topic and focus, and additionally stylistic aspects of the construction such as register […].
>
> *Corollary A*: If two constructions are syntactically distinct and Semantically synonymous, then they must not be Pragmatically synonymous.
>
> *Corollary B*: If two constructions are syntactically distinct and P[ragmatically] synonymous, then they must not be S[emantically] synonymous. (Goldberg 1995: 67)

This Principle of No Synonymy amplifies Haiman's Principle of Isomorphism (Haiman 1980: 516: "the commonly accepted axiom that no true synonyms exist, i.e., that different forms must have different meaning", citing linguists like Bloomfield and Bolinger), and largely assumes a structuralist view of language. Goldberg treats semantic or pragmatic distinctions as if they were analogous to distinctive features of phonemes (see, especially, Goldberg 2019: 23). Recent work points out, correctly, that Goldberg ignores the social role of variation. Leclerq & Morin (2023) instead suggest a Principle of No Equivalence:

> The *Principle of No Equivalence*: If two competing constructions differ in form (i.e. phonologically, morpho-syntactically or even orthographically), they must be semantically, pragmatically and/or socially distinct. (Leclercq & Morin 2023: 10)

While improving on Goldberg's proposal, the Principle of No Equivalence still does not account for patterns of semantic/functional equivalence without pragmatic differentiation or that are below the level of conscious social evaluation (i.e., Labov's [1972] Principle II), as in (6).

(6) Infinitive vs. gerund complementation:
She started **to eat** the apple vs. She started **eating** the apple. (Mair 2002)

Uhrig (2015) and others call these principles overrated given their obvious contradictions to the large body of variationist research (among other reasons). If contexts exist in which speakers can choose between alternative ways to express the same message, then the semantic boundaries between those formal structures must be at a minimum fuzzy. This notion of vagueness between forms (i.e., synonymy) is noted by linguists such as Erdmann (1910) and is inherent in the prototypicality effects that echo through cognitive linguistics and constructionist approaches. We additionally note that this issue has attracted considerable debate in sociolinguistics since at least the 1970s (see, e.g., Lavandera 1978; Labov 1978), with vagueness constituting an important pillar of third wave variationism (Eckert 2012; 2016; Moore 2021), in which variation is seen as inherently socially meaningful.

In an idealized description of constructional forms and their corresponding meaning, use, and/or language-external indexicality, one can suppose distinctions. For example, the dative alternation, as in (7) and (8), is traditionally explained by two different conceptualizations: a change of state construal (possession) favours the ditransitive, and a change of place construal (movement towards a goal) favours the prepositional variant (Gropen et al. 1989). However, this distinction only emerges from an empirical paradigm based on a grammarian's intuition or solicited grammaticality judgments. When tested against actual usage, this idealized explanatory factor has weak predictive power. Simply put, any semantic or functional difference between the ditransitive (7) and the prepositional (8) dative construction is neutralized in natural discourse. The traditional account (change of state or change of place) does not predict the choice as well as contextual constraints, such as pronominality of theme and recipient, definiteness of theme and recipient, length of theme and recipient, animacy of recipient, concreteness of theme, persistence etc. (Bresnan et al. 2007; Szmrecsanyi et al. 2017; Szmrecsanyi & Grafmiller 2023), indicating that a probabilistic account that involves constraints explains actual linguistic behaviour more accurately.

(7) Ditransitive dative: I gave give $[\text{him}]_{\text{recipient}}$ $[\text{some pizza.}]_{\text{theme}}$

(8) Prepositional dative: I gave [some pizza]$_{\text{theme}}$ to [him.]$_{\text{recipient}}$
(Szmrecsanyi et al. 2017)

Optionality is intrinsically sub-optimal for a theory of grammar that expects aesthetic or ontological parsimony/symmetry (Gunitsky 2019). Optionality inevitably triggers a certain type of complexity (Ma, Van Hoey & Szmrecsanyi 2025; Szmrecsanyi et al., in print; Van Hoey et al., accepted). If language A has multiple ways for expressing the same meaning and language B has only a single, isomorphic way, then language A is more complex than language B, i.e., language A's grammar requires a longer description to include all options. This is known as absolute complexity in the language complexity literature (see Miestamo 2008 for discussion). Complexity, however, can also be understood in a relative sense as the cost and difficulty for language users (again, see Miestamo 2008; see also Kusters 2003). For example, by definition, having two dative structures makes the English language more complex absolutely (i.e., there are more structure types to acquire). Implementing the probabilistic constraints that govern variant choice between dative options logically must also have additional cognitive costs for speakers of English relative to invariant structures.

Consider the *want to / wanna* alternation (*I want to go* vs. *I wanna go*), recently modelled and related to communicative efficiency by Levshina & Lorenz (2022). It is shown that these two variants are often not perceived as mutually exclusive alternatives targeting the formal-informal dichotomy. The distinction between them is regularly neutralized, and the variants are used interchangeably. Variant selection is subject to immediate contextual (not situational) constraints. That being said, upon reflection speaker may mentally classify the variants as formal/informal alternatives and then, according to Levshina and Lorenz, the Principle of No Synonymy will pull the variants apart functionally, creating a functional paradigm such that only one variant is grammatically licit in formal contexts while only the other is only grammatically licit in informal contexts. Even if full functional partitioning does not develop (i.e., heterogeneity persists), a variational probabilistic view of grammar in which social, contextual, language-internal, or any other influencing factor nudges speakers toward certain ways of speaking does suggest a certain validity to the Principle of No Synonymy, but as a pressure on a linguistic system rather than a as feature of it. Yet the theoretical ontological parsimony underlying the Principle of No

Synonymy itself is not evidenced. Diachronic trends towards isomorphism are still not proof of essential isomorphism.

Efficiency, in a broad sense, also relates to the relationship between absolute and relative complexity There is a widespread suspicion that absolute language complexity is positively correlated with relative complexity (Miestamo 2008; 2017). Miestamo further relates this predicted positive correlation between relative and absolute complexity to typological frequency and rarity:

> [I]f we take an inductive approach and look at what is actually found in the world's languages, and then try to evaluate the absolute complexity of those structures, it is highly likely that there will be some correlation between absolute complexity and cross-linguistic rarity. Perhaps (absolute) simplicity does not always mean ease of processing, but surely (absolute) complexity does in many cases add to processing difficulty. (Miestamo 2008: 38).

Miestamo (2008) predicts that longer absolute grammars will be more difficult to use and by extrapolation harder to acquire. Yet, this predicted positive correlation may be empirically elusive. Measures of absolute and relative complexity often do not capture it or, if they do, they overlook alternative explanations such as contextual specificity requirements (Hawkins 2019), register effects (Levshina & Lorenz 2022), or online processing and inference (Blumenthal-Dramé 2021), see overview by Ehret et al. (2021: 11–12). In summary, whether there is a positive correlation between absolute complexity and relative complexity remains an open empirical question, but the hypothesis is that the extra cognitive work required to handle absolutely complex grammar, regardless of how automatic and entrenched it is, results in increased cognitive load. This is on top of any other social, stylistic, or other considerations speakers are faced with when choosing one variant over another.

Against this backdrop, this paper investigates the extent to which grammatical optionality triggers production difficulty in a large corpus of naturalistic speech. The Principle of No Synonymy (and similar principles) predict such a triggering effect: if language(s) disfavour synonymy and optionality, then synonymy and optionality should be measurably sub-optimal in production. Likewise, optionality increases absolute complexity, and customary thinking about language complexity predicts that those increases in absolute complexity are paralleled by proportional increases in relative complexity (i.e. cost and difficulty). Whether all of this is empirically true is the question we address in this paper.

We thus examine 22 different grammatical variable contexts attested across varieties of spoken American English using the SWITCHBOARD corpus (Godfrey, Holliman & McDaniel 1992). Production difficulty (i.e. relative complexity) is assessed by locating and quantifying two different dysfluency types: (a) filled pauses, or delay markers; (b) unfilled pauses, or speech planning time (Ferreira 1991; Shriberg 1994; Berthold 1998; Clark & Wasow 1998; Berthold & Jameson 1999; Oomen & Postma 2001; Abel 2015; Lickley 2015; Le Grézause 2017). Hesitation phenomena, like speech planning time, have been used previously as metrics of relative cognitive effort (summarized by Berthold 1998; Berthold & Jameson 1999) and have been attested as more frequent in contexts independently judged to be more difficult, such as when utterances are longer or more syntactically complex (Grosjean, Grosjean & Lane 1979; Cooper & Paccia-Cooper 1980; Ferreira 1991; Shriberg 1994; Oviatt 1995; Clark & Wasow 1998; Lickley 2015; Christodoulides 2016), when the topic of conversation is unfamiliar (Smith & Clark 1993; Merlo & Mansur 2004), when the discursive task is more challenging (Oomen & Postma 2001; Abel 2015; Freeman 2015; Le Grézause 2017), or when lexical items are low frequency and/or have low contextual probability (Lieberman 1963; Tannenbaum & Williams 1968; see also Tily et al. 2009).

We replicate here the study by Gardner et al. (2021), who report, for a subset of the SWITCHBOARD corpus, that there is no significant (positive) relation between variation/optionality and dysfluency. In the current study, we expand upon Gardner et al. (2021) by extending the scope of the data to the full SWITCHBOARD corpus and by additionally classifying types of variation contexts/alternations following De Troij (2022).

This paper is structured as follows. We first present the methods (materials in Section 2.1; operationalization of variation contexts in Section 2.2; analytical framework in Section 2.3), followed by the results in Section 3.1; influence of alternation typology in Section 3.2, and finally a discussion and conclusion (Section 4).

2. Methods

The methodology of this study consists of four steps:

1. Tap into the SWITCHBOARD Corpus of spoken American English. As data points we consider each conversation. These materials are presented in Section 2.1.

2. Identify and quantify filled and unfilled pauses, the proxies for relative complexity (Section 2.1).

3. Identify and quantify variable contexts, i.e., sites where speakers have the choice between different grammatical ways of expressing the same meaning. In Section 2.2, we present the list of 22 commonly studied alternations, and present summary statistics.

4. Check if the there is a statistical correlation (Section 2.3).

2.1. Materials and (un)filled pauses

This paper analyses the full SWITCHBOARD corpus of spoken American English (Godfrey, Holliman & McDaniel 1992), an influential spoken corpus that consists of 2,438 spontaneous telephone conversations between 542 American English speakers who, in principle, are strangers to each other. The data were recorded by Texas Instruments in 1989–1990. Most recordings last 5 minutes, totalling to 240 hours for the whole SWITCHBOARD corpus. Demographic information about participants' age (15-69 years old), dialectal region, sex, and education level is available (Table 1), in addition to timestamps, previous places of residence etc. (Godfrey, Holliman & McDaniel 1992). While the connection between variation, dysfluency, and socio-demographic differences warrants investigation, we leave this topic to future publications.

Dialect group	**Age range**	**Sex**	**Education level**
South Midland (156)	15-19 (23)	Male (280)	Some (468)
Western (82)	20-29 (185)	Female (240)	None (52)
North Midland (75)	30-39 (146)		
Northern (73)	40-49 (108)		
Southern (55)	50-59 (54)		
New York City (31)	60+ (4)		
Mixed / Unknown (27)			
New England (21)			

Table 1: Demographics of the SWITCHBOARD corpus. Numbers in parentheses indicate number of individual speakers. Education level collapsed to some vs. no post-secondary education.

The public distribution of SWITCHBOARD includes time-aligned transcripts for the entire corpus. Additionally, previous work has already examined overt dysfluencies in portions of the corpus (e.g., Shriberg 1996; Clark & Fox Tree 2002; Wieling et al.

2016; Le Grézause 2017). Following Gardner et al. (2021), we consider all turn-internal uses of *um* and *uh* as overt hesitation markers, or FILLED PAUSES. This does not include other similar sounding terms like *um-hmm* or *uh-oh* or *um* and *uh* used as backchannels or turn initiators/enders, nor filled pauses at the start or end of a turn, because such filled pauses are most likely to be used to hold the floor. Further, we assume all remaining instances of *um* and *uh* are hesitation markers. This contrasts with analyses that treat them as tools for discourse organization (Clark & Fox Tree 2002) or dramatic performance. We recognize that this may be a limitation of our analysis; however, we point to the prolific literature showing higher rates of non-backchannel *um* and *uh* (regardless of function) when a speaker is objectively experiencing extra cognitive load (see Section 1). While Gardner et al. (2021) restrict their sample to a subset of young (20-29 years old) women who belong to the South Midland dialect region (285 conversations, 34 speakers, 7,161 turns), here we report on the full SWITCHBOARD corpus, in which after exclusions we find 58,032 filled pauses (*uh* and *um*) across the 2,438 conversations.

In addition to filled pauses, we measure speech planning time, or UNFILLED PAUSES. Maclay and Osgood's (1959) classic study identifies unfilled pauses as one of four major hesitation types (along with filled pauses, repeats, and false starts).[1] These hesitation types have been linked to production difficulties such as content planning, word retrieval, and the formulation of phrasal structure (Fox Tree & Clark 1997). Maclay & Osgood (1959) subjectively distinguish between short rhetorical pauses and unfilled pauses, which "were marked when there was judged to be an abnormal hesitation in speech […]" (Maclay & Osgood 1959: 24). We, on the other hand, follow Hieke et al. (1983), who identify 130 ms as the threshold beyond which pausing becomes psychologically rather than articulatorily motivated. We automatically identify unfilled pauses using the in-built script in the phonetics software program Praat (Boersma & Weenink 2023), called "Sound: To TextGrid (silences)". The main function of this script is to detect silence intervals in audio streams. We specify that silence is any part of the audio stream below -50 dB and longer than 130 ms. We also only consider turn-internal pausing and exclude pausing before or after speaking, or

[1]We are currently conducting a follow-up study of the SWITCHBOARD corpus that annotates false starts, repeats, and discourse markers in detail as well. As of December 2025, about the 50% of the audio files have additionally annotated. Preliminary analyses (some reported by Gardner & Szmrecsanyi 2022), find no positive correlation between choice points and restarts or repairs, and a very weak positive correlations for some discourse markers, like *you know*.

while an interlocutor is speaking. The total amount of turn-internal unfilled pause time is just over 61 hours within the 200+ hours of speech recording.

Both filled and unfilled pauses were standardized following the procedure laid out in Gardner et al. (2021). That is, relative frequencies were obtained by calculating the proportion of (un)filled pauses per 100 words, a standard procedure used when comparing speech samples that differ in length (as each conversation within SWITCHBOARD does).

2.2. Grammatical variation sites

The 22 grammatical alternations (i.e., variables) are the "usual suspects" in the literature on grammatical variation in American English (and beyond). When selecting the alternations we aimed to strike a balance between diachronic longevity (older and stable vs. newer), formality (no register association vs. prescriptive vs. informal speech), and grammaticalization status (wholly grammaticalized alternations vs. those involving a moderate degree of lexical material). The alternations subject to study can be categorized as involving (a) permutation, as in Table 2, (b) insertion / deletion, as in Table 3, or (c) substitution, as in Table 4 (De Troij 2022; Szmrecsanyi & Grafmiller 2023: 18–19). Permutation alternations exhibit a difference in the relative order of the items involved, e.g., *The man cut off his beard* vs *The man cut his beard off*. Insertion / deletion alternations are those in which a variant has one or more extra elements compared to the other one(s), e.g., *She know's* (that) *I'm coming on Friday*. Substitution are those cases in which variants differ with regard to the lexicalization of a single functional category, e.g., *A girl* has to / must *eat*. This typology is a refinement of Gries's (2017) overview article, in which he recognizes "word/constituent alternations" (permutation here), and "realization alternations" (insertion / deletion here). Yet, as the examples in Table 2-Table 4 show, most alternations studied by Gardner et al. (2021) originally, and in this current expanded replication, belong to the substitution type (N = 17), while the number of permutation (N = 3) and insertion / deletion (N = 2) type alternations is much smaller. We refer the reader to Gardner et al. (2021) for detailed descriptions and references to other studies of these 22 alternations.

In total, we identified 81,493 grammatical variable contexts in the SWITCHBOARD materials. These contexts were manually identified and carefully screened such that, in line with the variationist methodology (e.g., Tagliamonte 2012), substitution of an alternate variant would result in zero semantic/functional change in the specific

utterance that it occurred. To exemplify, an *s*-genitive variable context such as *my brother's sister* was licit because an *of*-genitive (*the sister of my brother*) would be semantically and functionally equivalent. Conversely, the *of*-genitive *three glasses of wine* was not licit as the corresponding *s*-genitive (**wine's three glasses*) would not be semantically or functionally equivalent. Additionally, invariant proper names like *Macy's* or *the United States of America* were excluded. Our specific criteria for inclusion/exclusion were based on the relevant variationist literature for each variable (again, see Gardner et al. 2021).

Permutation Alternation	**Example**	**Instances**
Particle placement	*He picked up the book.* *He picked the book up.*	3,272
Dative alternation	*The mother gave the baby the toy.* *The mother gave the toy to the baby.*	873
Genitive alternation	*the officer's uniform* *the uniform of the officer*	1,675
	TOTAL	5,820

Table 2: Examples of alternations that consist of permutations (N = 3).

Insertion /Deletion Alternation	**Example**	**Instances**
Complementizer	*I think that I know him.* *I think ∅ I know him.*	19,111
Relativization (restricted)	*The results that I obtained* *The results ∅ I obtained* *The results which I obtained*	944
	TOTAL	20,055

Table 3: Examples of alternations that consist of insertion and deletion (N = 2).

Substitution Alternation	**Example**	**Instances**
Comparatives (analytic vs. synthetic)	*Blood is thicker than water.* *Blood is more thick than water.*	766
Complementation (infinitive vs. gerund)	*I start to sing.* *I start singing.*	205
Complementation (*that* vs. gerund)	*I don't regret helping her.* *I don't regret that I helped her.*	307

Substitution Alternation	**Example**	**Instances**
Future marker	*I will leave tomorrow.* *I shall leave tomorrow.* *I am going to leave tomorrow* (etc.)	13,110
Deontic modality	*I must leave.* *I have to leave.* *I (have) got to leave.* *I need to leave.* (etc.)	7,998
Stative possession	*I have a headache.* *I (have) got a headache.*	5,576
Negation	*I do not want any.* *I do not want none.* *I want none.*	3,876
Not / auxiliary contraction	*She's not a student.* *She isn't a student.*	10,123
Indefinite pronouns	*Anybody home?* *Anyone home?*	5559
Coordinate pronouns	*He and I are going to the store.* *Him and me are going to the store.* *I and him are going to the store.* (etc.)	589
Quotatives	*And I went "Whoa!"* *And I was like "Whoa!"* *And I said, "Whoa!"* (etc.)	647
Try and/to/-ing	*I try to exercise daily.* *I try and exercise daily.* *I try exercising daily.*	1,124
Tried to/-ing	*I tried to exercise daily.* *I tried exercising daily.*	262
Without (any) / with no	*I ate salad without dressing.* *I ate salad with no dressing.*	82
Relativization (unrestricted)	*The shop that is new is great.* *The shop which is new is great.* (etc.)	763
There is/are plurals	*There are some erasers here.* *There is some erasers here.*	209
Verba dicendi *about/of*	*I think about the good old days.* *I think of the good old days.*	4,422
	TOTAL	55,618

Table 4: Examples of alternations that consist of substitutions (N = 17).

2.3. Quantitative analysis

After identifying and quantifying the two dependent variables – filled pauses (*uh/um*) and unfilled pauses (speech planning time) – and the predictor (relative frequency of variable contexts), we investigated the relation between each dependent variable and the predictor for both speakers in each conversation. Gardner et al. (2021) already determined that, for young Midland female speakers in SWITCHBOARD, there is no statistically significant positive relationship between dysfluency and variable contexts. This was true for both filled and unfilled pauses. Gardner et al. (2021) found that only their control variables of turn duration, mean word length, and speech rate were significant predictors of dysfluency.

Gardner et al. (2021) investigated the possible effect of individual grammatical alternations on dysfluency at the level of the conversational turn in a highly circumscribed group of SWITCHBOARD participants; in this paper we examine broad categories of alternations across the entire SWITCHBOARD data set. We take as our unit of analysis each dyadic conversation and ask the following: are conversations that are comparatively rich in variable contexts also more dysfluent compared to conversations that (for whatever reason) are less rich in variable contexts? The Principle of No Synonymy and other such doctrines of form-function symmetry, as Poplack & Dion 2009 describe them, predict that this should indeed be the case. To answer this question, we employ repeated measures correlation tests. Standard correlation tests (e.g., Pearson's correlation) assume that each data point is independent of all others. Most speakers in the SWITCHBOARD corpus took part in multiple conversations, so our per-speaker per-conversation data points are not fully independent. Therefore, we employ a statistical technique that accounts for any bias introduced by having multiple measures from the same participant (Bland & Altman 1995; Bakdash & Marusich 2017). Repeated measures correlations are calculated with the package *rmcorr* (Bakdash & Marusich 2022) in R (R Core Team 2023). The materials and analysis can be found in our accompanying Open Data Science (OSF) repository (see Data Availability Statement below).

3. Results

3.1. Do grammatical variable contexts in general predict dysfluencies?

In the 2,433 conversation transcripts from SWITCHBOARD there was an average of 33.48 variable contexts per conversation ($SD = 19.85$, $min = 1$, $max = 137$, $N_{total} = 81, 493$).

In total, we identified 58,032 *um*'s and *uh*'s (filled pauses), and about 3,788 minutes of turn-internal speech planning time (unfilled pauses). The number of filled pauses and unfilled pauses were converted into relative frequencies per hundred words to control for the differing lengths of conversations in SWITCHBOARD. Individual measures were recorded for both speakers in a conversation based only on their own speech. Next, we conducted two repeated measures correlations: first between filled pauses and variable contexts (Figure 1), then between unfilled pauses and variable contexts (Figure 2). Varying intercepts were calculated per speaker and are plotted in Figures 1 and 2.

Figure 1 shows the correlation between filled pauses (overt dysfluencies) and variable contexts per 100 words, while Figure 2 shows the correlation between unfilled pauses (speech planning time) and variable contexts per 100 words. Blue dots represent values for each speaker in individual conversations; thin blue lines represent regression lines for individual speakers across their multiple conversations. The black line represents the overall regression line.

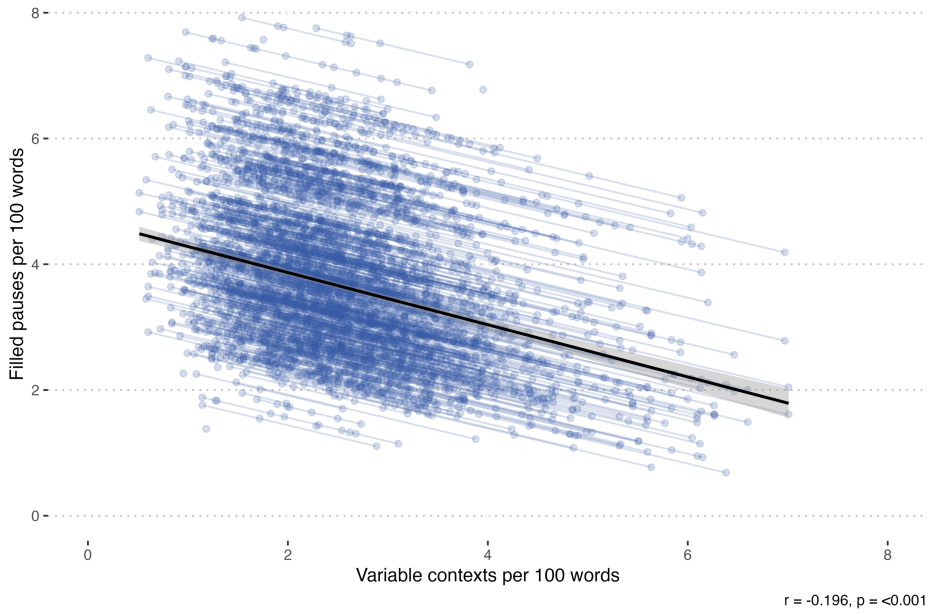


Figure 1: Repeated measures correlation between variable contexts and **filled** pauses ($N_{conversation}$ = 2,392, $N_{speaker}$ = 4,784). In this varying-intercept model, dots represent individual measurements per transcript. Thin lines represent regression lines per speaker. The black solid line represents the overall regression line. The slope (r = -0.20) indicates a weak negative correlation.

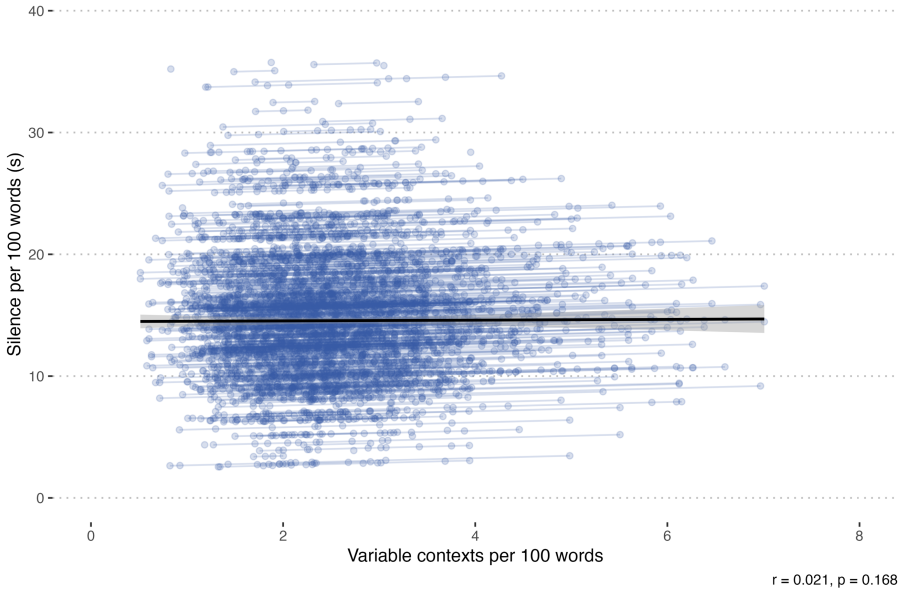


Figure 2: Repeated measures correlation between variable contexts and **unfilled** pauses ($N_{conversation} = 2392$, $N_{speaker} = 4784$). In this varying-intercept model, dots represent individual measurements per transcript. Thin lines represent regression lines per speaker. The black solid line represents the overall regression line. The slope ($r = 0.02$) indicates a very weak positive/no correlation.

As Figure 1 shows, the correlation between filled pauses and variable contexts is negative and significant, though weak ($r = -0.20$, $p < 0.001$). Correlation *r* values range from –1 (perfect negative correlation, or an increase in the x-axis independent variable predicts an equivalent decrease in the y-axis dependent variable) to +1 (perfect positive correlation, or an increase in the x-axis independent variable predicts an equivalent increase in the y-axis dependent variable); an *r* value of zero (flat regression line) indicates no relationship whatsoever (a change in the x axis variable does not predict any change in the y-axis variable). The correlation between unfilled pauses and variable contexts in Figure 2 is positive but extremely weak ($r = +0.02$); the relationship is not statistically significantly different from zero relationship ($p = 0.17$).

We note in this context that previous research (Le Grézause 2017) has suggested that there is a dyadic gender difference in the rates of filled pauses (mixed-gender dyads have more *um*s than same-gender dyads; same-gender dyads have more *uh*s than mixed-gender dyads). Investigating demographic meta-information, we can report that this is not borne out by our data (see the supplementary materials in the OSF repository).

In summary, we replicate the results reported in Gardner et al. (2021), using a much larger data set and by taking conversation (rather than turn) as the unit of interest. Overall, dysfluencies in the SWITCHBOARD simply do not correlate with grammatical optionality. However, the foregoing analysis does not distinguish between different types of alternations. This is the line of inquiry that we turn to next.

3.2. Does alternation type make a difference?

Not all variables involve the same sort of grammatical alternation. For example, insertion/deletion variables may facilitate Uniform Information Density (UID) optimization through reduction (Meister et al. 2021). Permutation variables are intimately linked with Easy First optimization (MacDonald 2013). Substitution variables often involve constraints that have to do with phonological or rhythmic well-formedness (Shih et al. 2015), and so on. It does make sense to conduct an additional series of repeated measures correlations (Figures 3–4) in which we split the data based on the nature of the alternation (see Tables 2–4).

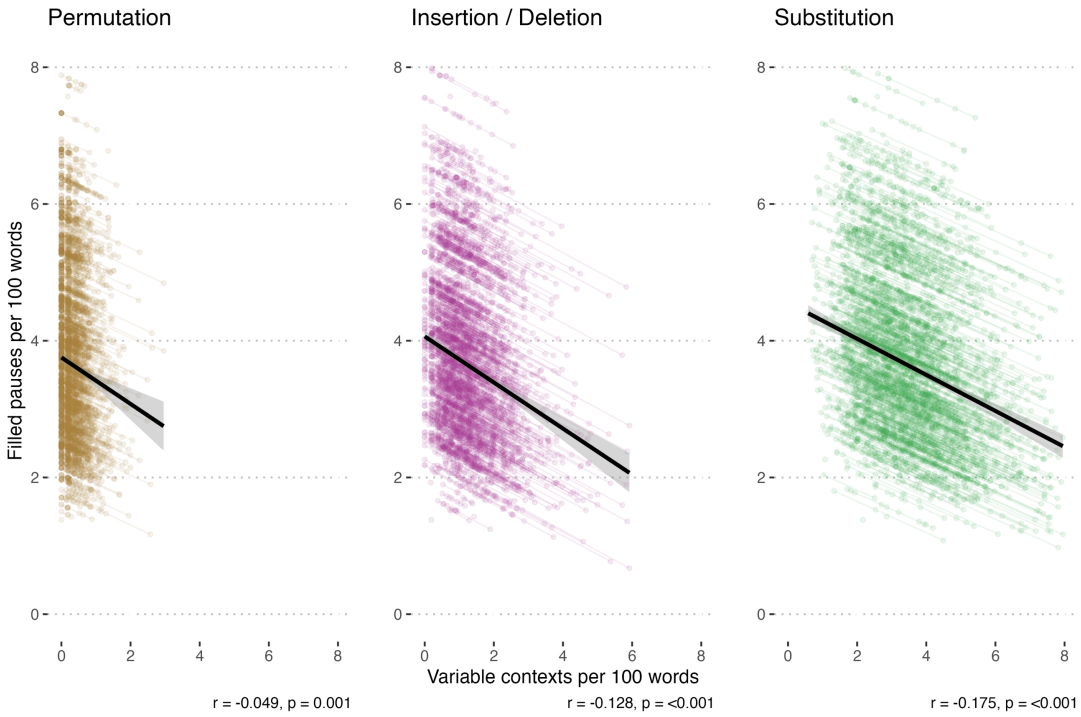


Figure 3: Repeated measures correlation for filled pauses, divided by subsets based on the three types of alternation ($N_{conversation}$ = 2392, $N_{speaker}$ = 4784). In this varying-intercept model, represent individual measurements per transcript. Thin lines represent regression lines per speaker. The black solid line represents the overall regression line.

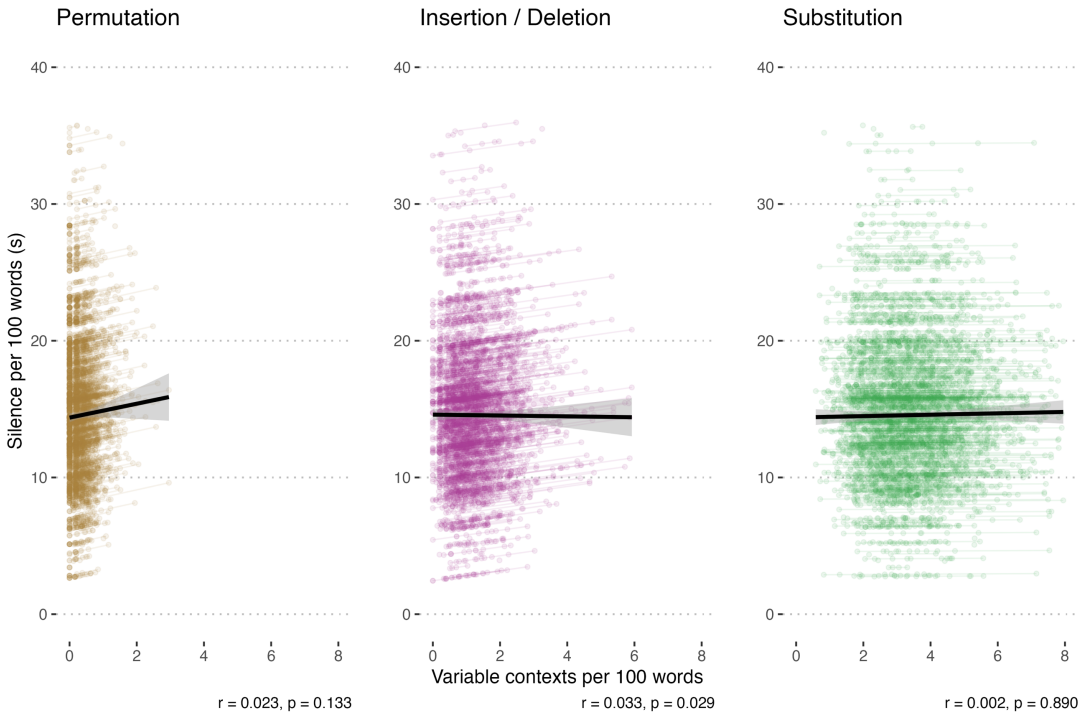


Figure 4: Repeated measures correlation for unfilled pauses, divided by subsets based on the three types of alternation ($N_{conversation} = 2392$, $N_{speaker} = 4784$). In this varying-intercept model, dots represent individual measurements per transcript. Thin lines represent regression lines per speaker. The black solid line represents the overall regression line.

Figure 3 shows that all correlations for filled pauses are negative and statistically significant ($p < 0.001$), meaning that each of the three categories of alternations repel dysfluency. The strongest of these negative correlation effects is for substitution alternations ($r = -0.18$), followed by insertion/deletion ($r = -0.13$) and permutation alternations ($r = -0.05$). Note that r values below –0.25 are considered weak, while those less than 0.10 are considered very weak. For unfilled pauses (Figure 4), only permutation alternations show a significant relationship between pausing and number of variable contexts ($r = 0.04$, $p = 0.007$). Both insertion/deletion and permutation alternations show a relationship that is not statistically significantly different from nil. This is congruous with our earlier finding that the number variable contexts per 100 words does not predict seconds of silence per 100 words (as in Figure 2).

As an interim summary, we note that distinguishing between different alternation types does not change the general observation that increased variable contexts does not positively predict either an increased number of dysfluencies or additional speech

planning time. In fact, for the most part, more grammatical optionality correlates (albeit weakly) with fewer dysfluencies and less speech planning time – otherwise the relationship cannot be statistically verified as beyond chance variation. Along with the results of Gardner et al. (2021), our findings provide strong evidence that variation and optionality do not result in additional production difficulty. That is, increased absolute complexity does not result in increased relative complexity. We discuss the implications of this finding in Section 4.

4. Discussion and conclusion

As we have shown, the often-presupposed positive correlation between relative complexity and absolute complexity is not borne out in our large corpus of spoken English. Absolute complexity was operationalized by identifying grammatical variable contexts, i.e., sites where speakers had to make an online choice between alternative forms. These contexts belong to 22 grammatical alternations that are well-studied in the variationist literature, which shows that the choice of alternates is based on non-trivial contextually sensitive probabilistic conditioning. Relative complexity was operationalized in two manners: filled pauses or delay markers, e.g., *uh* and *um*; and unfilled pauses or speech planning time. It is evident from the repeated measures correlations that speakers facing more sites of grammatical choice making are not disadvantaged cognitively by having to make those choices.

Further, grouping variables by typological similarity reproduces similar results. It is not the case, then, that insertion/deletion choices involve more cognitive challenge than permutation or substitution type choices. This is particularly noteworthy given that other characteristics of syntactic constructions or lexical items (especially their frequency) *do* affect a speaker's fluency in production. For example, in SWITCHBOARD, dysfluencies are more likely prior to unexpected lexical items (i.e., with high surprisal) or in contexts of more or longer syntactic dependencies (Dammalapati, Rajkumar & Agarwal 2019). Speakers have also been shown to actively manage the complexity of the utterances they produce (see, e.g., Rezaii et al. 2022). Given this, it may be the case that choice making sites only surface in utterances that are, all things being equal, easier to produce (e.g., with fewer content words and common syntactic structures). Optionality does not add relative complexity, but rather only surfaces in contexts that are already relatively simple. This is particularly the

case for permutation contexts. The corollary of this observation is that while absolute complexity does not add to relative complexity, the full suite of grammatical rules (including optional constructions) constituent of absolute complexity surface most robustly in utterances that are otherwise the least relatively complex.

We started out by contrasting the variationist view of linguistic variation — that it is endemic and normal — with a view underlying much theorizing in typology, functional linguistics, cognitive linguistics, and construction grammar: variation is sub-optimal and anomalous. This latter view is inseparable from the continued proposal that syntactic synonymy simply does not, cannot, or should not exist, viz. the Principle of Isomorphism (Haiman 1980), Principle of No Synonymy (Goldberg 1995), and Principle of No Equivalence (Leclercq & Morin 2023). The existence of near-synonymy at different linguistic levels is rarely denied outright, however, there is an assumption that such form-function or form-meaning overlap is necessarily short-lived and largely accidental (see De Smet et al. 2018; De Smet 2019 for critical discussion). Conversely, a foundational observation of variationist linguistics is that there exist instances in which forms and functions/meanings of different constructions overlap considerably, if not entirely, often because of language-external forces (Labov 1978). Variationists, however, do take great pains to identify the limits of variation, i.e., carefully circumscribing variable contexts. For example, there are scenarios in which variation could occur, but does not (e.g., *-ing* varies between [-ɪn] and [-iŋ], but never for one-syllable words like *king*), and scenarios where diachronic variation resolves to functional partitioning or loss (e.g., the functional partitioning of the *go* quotative for non-lexical sounds or gestures in Australian Aboriginal English, Rodríguez Louro et al. 2023). In modeling the probabilistic nature of three alternations (genitive, dative, particle placement) in different World Englishes, Szmrecsanyi & Grafmiller (2023) spend roughly 10 pages describing cases that do not participate in the variable context, i.e., where an isomorphic form-functional niche has been found.

Most typologists, functionalist/cognitive linguists, and construction grammarians tend to focus on the ways in which variants differ. If variants do have semantic/functional overlap, then this overlap is viewed as sub-optimal. They expect that the linguistic system will rebalance itself by finding different functional niches

over time, thereby repairing accidental form-function asymmetries (see Figure 5 for a visual depiction).

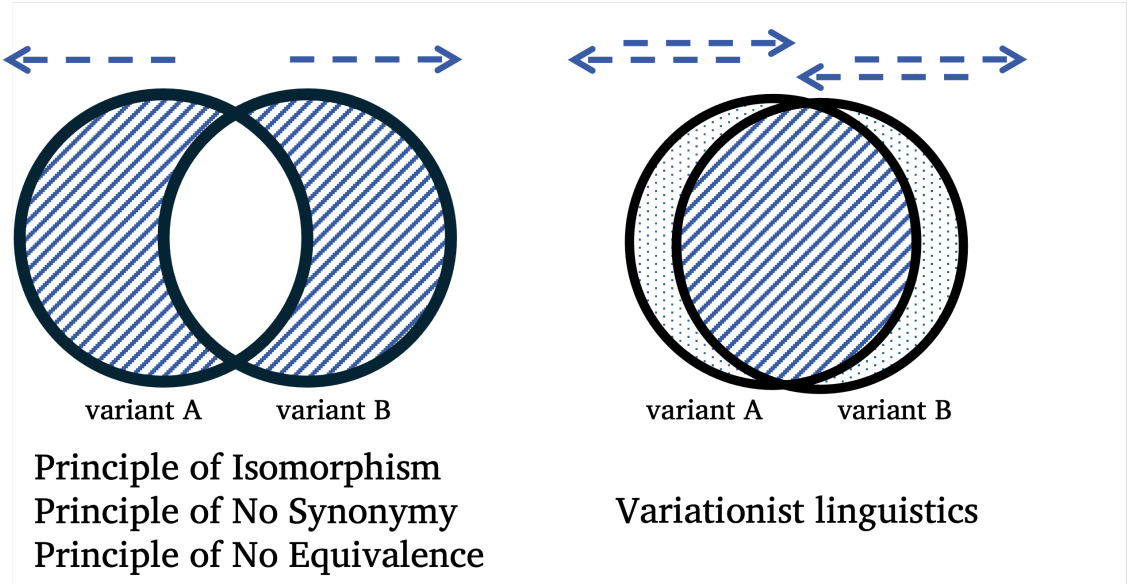


Figure 5. Diagrammatic representation of the Principle of Isomorphism / No Synonymy / No Equivalence (left) vs. Variationist Linguistics (right), their respective focus of attention (stripes), and view on diachronic change (arrows).

Poplack and Dion (2009: 557) call this way of thinking about optionality "the doctrine of form-function symmetry". Variationist linguists, by contrast, recognize that constructions can certainly have niche isomorphic functions or meanings, but their focus is on those instances where the boundaries between different forms and their linked meanings are indeed blurred. Consequently, the modeling of speakers' grammars will then necessarily involve the modeling of probabilistic grammars, which in essence are not unstable in nature but are dynamic, across time and situation (indicated by arrows in Figure 5). In this regard, grammatical optionality is not fundamentally problematic, but in fact a feature of language. Thus, whether variation is a blemish on the linguistic system or a fundamental part of it depends on which analytic glasses you wear (either the left or right side of Figure 5). Our findings, however, only evidence one of these views: variation is not a problem.

Indeed, there may even be cognitive benefits to having multiple variants available, as already argued in Gardner et al. (2021). For example, speakers can adjust their explicitness depending on the complexity of the environment (Rohdenburg 1996), manage information density (Levy & Jaeger 2007; Meister et al. 2021), be biased to produce easy constituents first or heavy constituents last ("Easy First" in MacDonald

2013; “End Weight” in Eitelmann 2016), or produce a more eurhythmic utterance (Shih et al. 2015). So perhaps having multiple grammatical ways of saying the same thing is a blessing rather than a blight, an issue that warrants further exploration by including other forms of dysfluencies, such as discourse markers, restarts and repairs.

This is the first study of a large representational corpus of a major variety of spoken English in which the notion of complexity was interpreted as absolute (operationalized as 22 major types of grammatical alternations) vs. relative (filled and unfilled pauses). The methodology should be applied to other spoken corpora. In particular, attention should be devoted to the investigation of other kinds of alternations, e.g., phonological or lexical, next to grammatical alternations (see examples 3-5 in the introduction). One may also wonder whether these findings are restricted to English data (unlikely) or also appear in other languages. To that end, we are currently investigating spoken data from Mandarin Chinese. While our study is corpus-based, additional psycholinguistic classification tasks provide potential points of comparison to cross-validate our results. And finally, there may be a difference in the difficulty of choices that need to be made. Some turns may contain contexts that are constrained (easier choices), other may be more restricted (harder choices). It is not unthinkable that hard choices would elicit more hesitation markers, i.e., come with a higher relative cost. We have modeled this aspect on the same dataset (Szmrecsanyi et al., in print; Van Hoey et al., accepted), with similar results as the current contribution. For now, however, it is clear that the predicted theoretical positive relation between relative and absolute complexity was not found in our data. Absolute complexity does not necessarily trigger relative complexity after all.

Data availability statement

Data and code can be found on the accompanying OSF repository.
https://osf.io/at5mn/

Acknowledgements

Funding by the KU Leuven Research Council (grant # 3H220293) is gratefully acknowledged. We also wish to thank the useful feedback received at the conference on “Naturally occurring data in and beyond linguistic typology” (2023, University of Bologna), as well as the editors for their support w.r.t. this special issue.

References

Abel, Jennifer Colleen. 2015. *The effect of task difficulty on speech convergence*. Doctoral dissertation, University of British Columbia.

Bakdash, Jonathan Z. & Laura R. Marusich. 2017. Repeated Measures Correlation. *Frontiers in Psychology* 8. 456. https://doi.org/10.3389/fpsyg.2017.00456.

Bakdash, Jonathan Z. & Laura R. Marusich. 2022. *rmcorr: Repeated measures correlation*. Manual.

Berthold, André. 1998. Repräsentation und Verarbeitung sprachlicher Indikatoren für kognitive Ressourcenbeschränkungen. MA disseration, Universität des Saarlandes.

Berthold, André & Anthony Jameson. 1999. Interpreting symptoms of cognitive load in speech input. In Judy Kay (ed.), *User Modeling: Proceedings of the Seventh International Conference, UM99*, 235–244. New York: Springer.

Bland, J Martin. & Douglas G. Altman. 1995. Statistics notes: Calculating correlation coefficients with repeated observations: Part 1--correlation within subjects. *BMJ* 310(6977). 446–446. https://doi.org/10.1136/bmj.310.6977.446.

Blumenthal-Dramé, Alice. 2021. The online processing of causal and concessive relations: Comparing native speakers of English and German. *Discourse Processes* 58(7). 642–661. https://doi.org/10.1080/0163853X.2020.1855693.

Boersma, Paul & David Weenink. 2023. Praat: Doing phonetics by computer. www.praat.org.

Bolinger, Dwight L. 1968. Entailment and the meaning of structures. *Glossa* 2. 119–127.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina & Harald Baayen. 2007. Predicting the dative alternation. In Gerlof Boume, Irene Krämer & Joost Zwarts (eds.), *Cognitive foundations of interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Bresnan, Joan & Marilyn Ford. 2010. Predicting syntax. *Language* 86(1). 168–213. https://doi.org/10.1353/lan.0.0189.

Christodoulides, George. 2016. *Effects of cognitive load on speech production and perception*. Doctoral dissertation, Université catholique de Louvain, Louvain-la-Neuve.

Clark, Eve V. 1987. The principle of contrast: A constraint on language acquisition. In Brian MacWhinney (ed.), *Mechanisms of language acquisition*, 1–33. Hillsdale, NJ: Lawrence Erlbaum Associates.

Clark, Herbert H. & Jean E. Fox Tree. 2002. Using *uh* and *um* in spontaneous speaking. *Cognition* 84(1). 73–111. https://doi.org/10.1016/S0010-0277(02)00017-3.

Clark, Herbert H. & Thomas Wasow. 1998. Repeating words in spontaneous speech. *Cognitive Psychology*. Elsevier BV 37(3). 201–242. https://doi.org/10.1006/cogp.1998.0693.

Cooper, William E. & Jeanne Paccia-Cooper. 1980. *Syntax and speech*. *Syntax and Speech*. Cambridge, MA: Harvard University Press.

Dammalapati, Samvit, Rajakrishnan Rajkumar & Sumeet Agarwal. 2019. Expectation and locality effects in the prediction of disfluent fillers and repairs in English speech. In *American Chapter of the Association for Computational Linguistics: Student research workshop*, 103–109. Minneapolis, MN: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-3015.

De Smet, Hendrik. 2019. The motivated unmotivated: Variation, function and context. In Kristin Bech & Ruth Möhlig-Falke (eds.), *Grammar – Discourse – Context*, 305–332. Berlin: De Gruyter. https://doi.org/10.1515/9783110682564-011.

De Smet, Hendrik, Frauke D'hoedt, Lauren Fonteyn & Kristel Van Goethem. 2018. The changing functions of competing forms: Attraction and differentiation. *Cognitive Linguistics* 29(2). 197–234. https://doi.org/10.1515/cog-2016-0025.

De Troij, Robbert. 2022. *Natiolectal variation in Dutch grammar: A data-driven approach*. Doctoral dissertation, Radboud University, Nijmegen; KU Leuven.

Eckert, Penelope. 2008. Variation and the indexical field. *Journal of Sociolinguistics* 12(4). 453–76. https://doi.org/10.1111/j.1467-9841.2008.00374.x.

Eckert, Penelope. 2012. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology* 41. 87–100.

Eckert, Penelope. 2016. Variation, meaning and social change. In Nikolas Coupland (ed.), *Sociolinguistics: Theoretical debates*, 68–85. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781107449787.004.

Ehret, Katharina, Alice Blumenthal-Dramé, Christian Bentz & Aleksandrs Berdicevskis. 2021. Meaning and measures: Interpreting and evaluating complexity metrics. *Frontiers in Communication* 6. 640510. https://doi.org/10.3389/fcomm.2021.640510.

Eitelmann, Matthias. 2016. Support for end-weight as a determinant of linguistic variation and change. *English Language and Linguistics* 20(3). 395–420. https://doi.org/10.1017/S1360674316000356.

Erdmann, Karl-Otto. 1910. *Die Bedeutung des Wortes: Aufsätze aus dem Grenzgebiet der Sprachpsychologie und Logik*. 2nd edn. Leipzig: Avenarius.

Ferreira, Fernanda. 1991. Effects of length and syntactic complexity on initiation times for prepared utterances. *Journal of Memory and Language* 30(2). 210–233. https://doi.org/10.1016/0749-596x(91)90004-4.

Fox Tree, Jean E. & Herbert H Clark. 1997. Pronouncing "the" as "thee" to signal problems in speaking. *Cognition* 62(2). 151–167. https://doi.org/10.1016/S0010-0277(96)00781-0.

Freeman, Valerie. 2015. *The phonetics of stance-taking*. Doctoral dissertation, University of Washington.

Gardner, Matt Hunt & Benedikt Szmrecsanyi. 2022. *Um, uh, and variation in American English*. (Paper presented at Methods XVII - Methods in Dialectology & Language Diversity, Johannes Gutenberg-University, Mainz, Germany, August 1-5, 2022).

Gardner, Matt Hunt, Eva Uffing, Nicholas Van Vaeck & Benedikt Szmrecsanyi. 2021. Variation isn't that hard: Morphosyntactic choice does not predict production difficulty. (Ed.) Stefan Th. Gries. *PLOS ONE* 16(6). e0252602. https://doi.org/10.1371/journal.pone.0252602.

Godfrey, John J., Edward C. Holliman & Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 517–520. IEEE Computer Society. https://doi.org/10.1109/ICASSP.1992.225858

Goldberg, Adele E. 1995. *Constructions*. Chicago: University of Chicago Press.

Goldberg, Adele E. 2019. Explain me this: Creativity, competition, and the partial productivity of constructions. Princeton (N.J.): Princeton University Press.

Gries, Stefan Th. 2017. Syntactic alternation research: Taking stock and some suggestions for the future. *Belgian Journal of Linguistics* 31. 8–29. https://doi.org/10.1075/bjl.00001.gri.

Gropen, Jess, Steven Pinker, Michelle Hollander, Richard Goldberg & Ronald Wilson. 1989. The learnability and acquisition of the dative alternation in English. *Language* 65(2). 203. https://doi.org/10.2307/415332.

Grosjean, François, Lysiane Grosjean & Harlan Lane. 1979. The patterns of silence: Performance structures in sentence production. *Cognitive Psychology*. Elsevier BV 11(1). 58–81. https://doi.org/10.1016/0010-0285(79)90004-5.

Gunitsky, Seva. 2019. Rival visions of parsimony. *International Studies Quarterly* 63(3). 707–716. https://doi.org/10.1093/isq/sqz009.

Haiman, John. 1980. The iconicity of grammar: Isomorphism and motivation. *Language* 56(3). 515–540. https://doi.org/10.2307/414448.

Haiman, John. 1985. *Natural syntax: Iconicity and erosion*. Cambridge: Cambridge University Press.

Hawkins, John A. 2019. Word-external properties in a typology of Modern English: a comparison with German. *English Language and Linguistics* 23(3). 701–727. https://doi.org/10.1017/S1360674318000060.

Hieke, Adolf E., Sabine Kowal & Daniel C. O'Connell. 1983. The trouble with "articulatory" pauses. *Language and Speech* 26(3). 203–214. https://doi.org/10.1177/002383098302600302.

Hout, Roeland van & Pieter Muysken. 2016. Taming chaos. Chance and variability in the language sciences. In Klaas Landsman & Ellen van Wolde (eds.), *The Challenge of Chance*, 249–266. Cham: Springer. https://doi.org/10.1007/978-3-319-26300-7_14.

Kusters, Wouter. 2003. *Linguistic complexity: The influence of social change on verbal inflection* (LOT 77). Utrecht: LOT.

Labov, William. 1972a. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Labov, William. 1972b. Some principles of linguistic methodology. *Language in Society* 1. 97–120.

Labov, William. 1978. Where does the linguistic variable stop? A response to Batriz R. Lavandera. *Working Papers in Sociolinguistics* 44. 1–22.

Labov, William. 1984. Field methods of the project of linguistic change and variation. In John Baugh & Joel Sherzer (eds.), *Language in use. Readings in sociolinguistics*, 28–53. Englewood Cliffs, NJ: Prentice Hall.

Lavandera, Beatriz R. 1978. Where does the linguistic variable stop? *Language in Society* 7(2). 171–182.

Le Grézause, Esther. 2017. Um *and* uh*, and the expression of stance in conversational speech*. Doctoral dissertation, University of Washington.

Leclercq, Benoît & Cameron Morin. 2023. No equivalence: A new principle of no synonymy. *Constructions* 1–16. https://doi.org/10.24338/CONS-535.

Levshina, Natalia & David Lorenz. 2022. Communicative efficiency and the Principle of No Synonymy: Predictability effects and the variation of *want to* and *wanna*. *Language and Cognition* 14(2). 249–274. https://doi.org/10.1017/langcog.2022.7.

Levy, Roger & T Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. In Bernard Schölkopf, John Platt & Thomas Hoffman (eds.), *Advances in Neural Information Processing Systems 19*, 849–856. Cambridge, MA: MIT Press. http://books.nips.cc/papers/files/nips19/NIPS2006_0515.pdf.

Lickley, Robin J. 2015. Fluency and disfluency. In Melissa A. Redford (ed.), *The handbook of speech production*, 445–469. Oxford: Wiley-Blackwell.

Lieberman, Philip. 1963. Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech* 6(3). 172–187. https://doi.org/10.1177/002383096300600306.

Ma, Ruiming, Thomas Van Hoey & Benedikt Szmrecsanyi. 2025. Isomorphism-inspired theorising about optionality and variation: no empirical support from English grammar. *English Language and Linguistics*. https://doi.org/10.1017/S1360674325000097.

MacDonald, Maryellen C. 2013. How language production shapes language form and comprehension. *Frontiers in Psychology* 4. 1–16. https://doi.org/10/gbfpt3.

Maclay, Howard & Charles E. Osgood. 1959. Hesitation phenomena in spontaneous English speech. *WORD* 15(1). 19–44. https://doi.org/10.1080/00437956.1959.11659682.

MacWhinney, Brian. 1989. Competition and lexical categorization. In Roberta Corrigan, Fred Eckman & Michael Noonan (eds.), *Current issues in linguistic theory. Vol 61: Linguistic categorization* (Amsterdam Studies in the Theory and History of Linguistic Science 4), 195–241. Amsterdam: John Benjamins.

Mair, Christian. 2002. Three changing patterns of verb complementation in Late Modern English: A real-time study based on matching text corpora. *English Language and Linguistics* 6(1). 105–131. https://doi.org/10.1017/s1360674302001065.

Meister, Clara, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell & Roger Levy. 2021. Revisiting the Uniform Information Density Hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 963–980. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.74.

Merlo, Sandra & Letícia Lessa Mansur. 2004. Descriptive discourse. *Journal of Communication Disorders* 37(6). 489–503. https://doi.org/10.1016/j.jcomdis.2004.03.002.

Miestamo, Matti. 2008. Grammatical complexity in a cross-linguistic perspective. In Matti Miestamo, Kaius Sinnemäki & Fred Karlsson (eds.), *Language complexity: typology, contact, change*, 23–42. Amsterdam, Philadelphia: Benjamins.

Miestamo, Matti. 2017. Linguistic diversity and complexity. *Lingue e Linguaggio* 16(2). 227–253.

Moore, Emma. 2021. The social meaning of syntax. In Lauren Hall-Lew, Emma Moore & Robert J. Podesva (eds.), *Social meaning and linguistic variation*, 54–79. 1st edn. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108578684.003.

Oomen, Claudy & Albert Postma. 2001. Effects of divided attention on the production of filled pauses and repetitions. *Journal of Speech, Language, and Hearing Research* 44. 997–1004. https://doi.org/10.1044/1092-4388(2001/078).

Oviatt, Sharon. 1995. Predicting spoken disfluencies during human–computer interaction. *Computer Speech & Language* 9(1). 19–35. https://doi.org/10.1006/csla.1995.0002.

Poplack, Shana & Nathalie Dion. 2009. Prescription vs praxis: The evolution of future temporal reference in French. *Language*. JSTOR 85(3). 557–587.

R Core Team. 2023. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rezaii, Neguine, Kyle Mahowald, Rachel Ryskin, Bradford Dickerson & Edward Gibson. 2022. A syntax–lexicon trade-off in language production. *Proceedings of the National Academy of Sciences* 119(25). e2120203119. https://doi.org/10.1073/pnas.2120203119.

Rodríguez Louro, Celeste, Glenys Dale Collard, Madeleine Clews & Matt Hunt Gardner. 2023. Quotation in earlier and contemporary Australian Aboriginal English. *Language Variation and Change* 35(2). 129–152. https://doi.org/10.1017/S0954394523000169.

Rohdenburg, Günter. 1996. Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics* 7(2). 149–182. https://doi.org/10.1515/cogl.1996.7.2.149.

Shih, Stephanie, Jason Grafmiller, Richard Futrell & Joan Bresnan. 2015. Rhythm's role in genitive construction choice in spoken English. In Ralf Vogel & Ruben Vijver (eds.), *Rhythm in cognition and grammar*. Berlin: De Gruyter. https://doi.org/10.1515/9783110378092.207.

Shriberg, Elizabeth. 1994. *Preliminaries to a theory of speech disfluencies*. Doctoral dissertation, University of California, Berkeley.

Shriberg, Elizabeth E. 1996. Disfluencies in Switchboard. In H. Timothy Bunnell & Richard A. Foulds (eds.), *International Conference on Spoken Language Processing, Addendum*. Wilmington, DW: Alfred I. duPont Institute.

Smith, Vicki L. & Herbert H. Clark. 1993. On the course of answering questions. *Journal of Memory and Language* 32(1). 25–38. https://doi.org/10.1006/jmla.1993.1002.

Suikkanen, Jussi. 2018. Deontic modality. *Analysis* 78(2). 354–363. https://doi.org/10.1093/analys/any015.

Szmrecsanyi, Benedikt. 2017. Variationist sociolinguistics and corpus-based variationist linguistics: Overlap and cross-pollination potential. *Canadian Journal of Linguistics/Revue canadienne de linguistique* 62(4). 685–701. https://doi.org/10.1017/cnj.2017.34.

Szmrecsanyi, Benedikt, Matt Hunt Gardner, Ruiming Ma & Thomas Van Hoey. in print. Empirical accountability meets theorizing about language variation. In Patricia Cukor-Avila, Sali A. Tagliamonte & Guy Bailey (eds.), *Empirical accountability in variation linguistics: Taking the next step*. Cambridge: Cambridge University Press.

Szmrecsanyi, Benedikt & Jason Grafmiller. 2023. *Comparative variation analysis: Grammatical alternations in world Englishes* (Studies in Language Variation and Change). Cambridge: Cambridge University Press.

Szmrecsanyi, Benedikt, Jason Grafmiller, Joan Bresnan, Anette Rosenbach, Sali Tagliamonte & Simon Todd. 2017. Spoken syntax in a comparative perspective: The dative and genitive alternation in varieties of English. *Glossa: a journal of general linguistics* 2(1). 86. https://doi.org/10.5334/gjgl.310.

Tagliamonte, Sali A. 2012. *Variationist sociolinguistics: Change, observation, interpretation* (Language in Society 40). Malden, MA: Wiley-Blackwell.

Tagliamonte, Sali A. & Katharina Pabst. 2020. A *cool* comparison: Adjectives of positive evaluation in Toronto, Canada and York, England. *Journal of English Linguistics* 48(1). 3–30. https://doi.org/10.1177/0075424219881487.

Tannenbaum, Percy H. & Frederick Williams. 1968. Generation of active and passive sentences as a function of subject or object focus. *Journal of Verbal Learning and Verbal Behavior* 7(1). 246–250. https://doi.org/10.1016/S0022-5371(68)80197-5.

Tily, Harry, Susanne Gahl, Inbal Arnon, Neal Snider, Anubha Kothari & Joan Bresnan. 2009. Syntactic probabilities affect pronunciation variation in spontaneous speech. *Language and Cognition* 1(2). 147–165. https://doi.org/10.1515/LANGCOG.2009.008.

Uhrig, Peter. 2015. Why the Principle of No Synonymy is overrated. *Zeitschrift für Anglistik und Amerikanistik; Leipzig*. Leipzig, Germany, Leipzig: Walter de Gruyter GmbH 63(3). 323–337. http://doi.org/10.1515/zaa-2015-0030.

Van Hoey, Thomas, Matt H. Gardner, Ruiming Ma & Benedikt Szmrecsanyi. Accepted. Unpredictable grammatical choices are not harder than predictable ones. *Language Variation and Change*.

Wälchli, Bernhard & Anna Sjöberg. 2025. A law of meaning. *Linguistic Typology at the Crossroads*. Linguistic Typology at the Crossroads 4(2). 1–71. https://doi.org/10.6092/ISSN.2785-0943/18920.

Wieling, Martijn, Jack Grieve, Gosse Bouma, Josef Fruehwald, John Coleman & Mark Liberman. 2016. Variation and change in the use of hesitation markers in Germanic languages. *Language Dynamics and Change* 6(2). 199–234. https://doi.org/10.1163/22105832-00602001.

Contact

thomas.vanhoey@kuleuven.be
benedikt.szmrecsanyi@kuleuven.be