Linguistic Typology at the Crossroads 3-2 (2023): 100-131 ISSN 2785-0943

Using a parallel corpus to study patterns of word order variation: determiners and quantifiers within the noun phrase in European languages

LUIGI TALAMO

LANGUAGE SCIENCE AND TECHNOLOGY, SAARLAND UNIVERSITY (GERMANY)

Submitted: 19/10/2022 Revised version: 21/06/2023
Accepted: 28/07/2023 Published: 27/12/2023

Articles are published under a Creative Commons Attribution 4.0 International License (The authors remain the copyright holders and grant third parties the right to use, reproduce, and share the article).

Abstract

Despite the wealth of studies on word order, there have been very few studies on the order of minor word categories such as determiners and quantifiers. This is likely due to the difficulty of formulating valid cross-linguistic definitions for these categories, which also appear problematic from a computational perspective. A solution lies in the formulation of comparative concepts and in their computational implementation by combining different layers of annotation with manually compiled list of lexemes; the proposed methodology is exemplified by a study on the position of these categories with respect to the nominal head, which is conducted on a parallel corpus of 17 European languages and uses Shannon's entropy to quantify word order variation. Whereas the entropy for the article-noun pattern is, as expected, extremely low, the proposed methodology sheds light on the variation of the demonstrative-noun and the quantifier-noun patterns in three languages of the sample.

Keywords: word order; determiner; quantifier; entropy; Universal Dependency; European languages

1. Introduction

Most of the previous studies on word order have been focused on major constituents like subject, verb and object, or adjectives and nouns. Although the two categories of demonstratives and numerals figure in many classic typologies on word order

(Greenberg 1963; Dryer 2009; Hawkins 1983), the closely related categories of articles and non-numerical quantifiers have received little attention (Ioup 1975; Greenberg 1978; Dryer 1992). Quantitative typological studies (Futrell et al. 2015; Naranjo & Becker 2018; Alzetta et al. 2018; Gerdes et al. 2019; Levshina 2019; Talamo & Verkerk 2022), which exploit computational resources, such as annotated treebanks and parsed corpora, and interpret the frequency through information-theoretic measures, have so far not considered these categories either.

The reason behind the neglect of these categories lies in the objective difficulty of defining determiners and non-numerical quantifiers. A quick look to grammars shows that demonstratives are often conflated together with other nominal modifiers such as articles and non-numerical quantifiers; whereas a category of numerical quantifiers, or 'numerals', can be quite easily identified, non-numerical quantifiers are often treated together with adjectives, numerals or even non-nominal modifiers like adverbs and intensifiers.

Previous qualitative studies that explicitly consider one of these categories employ a categorical measure to describe the word order pattern i.e., only one possible word pattern can be assigned to a language. On the other hand, quantitative studies use continuous measure such as frequency to capture the variability of word patterns, but the annotation schemata on which they are based do not offer fine-grained distinctions for determiners and non-numerical quantifiers.

In the present paper I aim to fill this gap by looking at the frequencies of noun-article, noun-demonstrative and noun-quantifier orders in a parallel corpus of 17 European languages, which is automatically parsed using tools from the Universal Dependency (UD) project. The rest of the paper is structured as follows: Sect. 2 briefly reviews the qualitative and quantitative studies on the order of determiners and quantifiers; Sect. 3 describes the methodology, presenting the parallel corpus, the information-theoretic measure used to interpret the frequencies and the implementation of the comparative concepts using annotations from the UD project; Sect. 4 presents the results and gives an in-depth analysis of a selection of word order patterns showing high variability; Sect. 5 concludes.

2. The order of determiners and quantifiers within the noun phrase in qualitative and quantitative studies

The term 'determiners' is widely employed as an umbrella term for both articles and demonstratives, which is problematic even for a small sample like the one used in the

present article. As already observed by Dryer (2007: 152, 161-162), there are languages in which articles are used in combination with demonstratives and other types of determiners, like possessives, and there are languages in which articles do not exist. In my sample, Greek (ell; Indo-European, Graeco-Phrygian), Irish (gle; Indo-European, Celtic) and Welsh (cym; Indo-European, Celtic) are languages of the former type, while Bosnian-Croatian-Serbian (BCS[1]; hbs; Indo-European, Balto-Slavic), Bulgarian (bul; Indo-European, Balto-Slavic), Czech (ces; Indo-European, Balto-Slavic), Lithuanian (lit; Indo-European, Balto-Slavic), Polish (pol; Indo-European, Balto-Slavic) and Russian (rus; Indo-European, Balto-Slavic) are languages of the latter type.

The term demonstrative is often used interchangeably for both stand-alone words i.e., demonstrative pronouns, and modifiers; the latter are further divided into nominal demonstratives and adverbial demonstratives, which are usually etymologically connected; cfr. English (eng; Indo-European, Germanic) *this* and *that* vs. *here* and *there* (Diessel & Coventry 2020: 1). I consider here nominal demonstratives, which are sometimes described by grammars as 'demonstrative adjectives' or 'adnominal demonstratives' (Verkerk, p.c.), and I refer here to them as 'demonstratives'.

As for non-numerical quantifiers, the term is often kept distinct from the similar category of numerical quantifiers, or 'numerals', indicating non-numerical words that express quantity; I refer here to this category as 'quantifiers'. The category of quantifiers is from time to time lumped with determiners and/or adjectives, as in the following quotation from a recent grammar of Irish:

> A variety of words referring to quantities also function as determiners within NPs. […] They are on the whole rather a mixed bag of elements from a syntactic point of view. Many of these forms are treated as adjectives in traditional grammars, although they cannot be declined or compared like the adjectives. […] (Stenson 2020: 188)

Studies on the order of articles and demonstratives with respect to the nominal head go back at least to Greenberg (1963), where Universal 18 is formulated as follows: "When the descriptive adjective precedes the noun, the demonstrative, and the

[1] I follow Alexander (2006)'s usage of the acronym BCS to indicate the pluricentric language formerly known as Serbo-Croatian.

numeral, with overwhelmingly more than chance frequency, does likewise." (Greenberg 1963: 68).

language	art & noun	dem & noun	quant & noun
BCS	-	dem-noun	quant-noun
Bulgarian	-	dem-noun	quant-noun
Czech	-	dem-noun	quant-noun
Danish	art-noun	dem-noun	quant-noun
Dutch	art-noun	dem-noun	quant-noun
English	art-noun	dem-noun	quant-noun
French	art-noun	dem-noun	quant-noun
German	art-noun	dem-noun	quant-noun
Greek	art-noun	dem-noun	quant-noun
Irish	art-noun#	noun-dem	quant-noun
Lithuanian	-	dem-noun	quant-noun
Polish	-	dem-noun	quant-noun
Portuguese	art-noun	dem-noun	quant-noun
Romanian	art-noun*	mixed	quant-noun
Russian	-	dem-noun	quant-noun
Spanish	art-noun	dem-noun	quant-noun
Welsh	art-noun#	noun-dem	quant-noun

Table 1: The order of articles, demonstratives and quantifiers with respect to the nominal head according to Dryer (2008, 2013a), Siewerska (1998). *: only indefinite articles; #: only definite articles.

The two categories of dependents also feature in subsequent studies such as Hawkins (1983) and Dryer (1992, 2009); according to Dryer (1992, 2009), articles and demonstratives figure among the categories of dependents that do not support the tendency for which dependents follow heads in VO languages and precede in OV languages. Rather than treating word order correlations as a "tendency towards consistent ordering of heads and dependents" (Dryer 1992: 82) as in the previous Head-Dependent Theory (HDT), Dryer's Branching-Direction Theory (BDT) postulates that constituents follow the same position of either the verb or the object in the verb-object ordering; in a sample of 675 languages, later expanded to over 1500 languages in his 2009 article, Dryer (1992) finds that articles follow the same position of verb i.e. are verb patterners, while demonstratives follow the same position of object i.e., are object patterners. This explains why, from the perspective of the HDT, articles

and demonstratives behave like heads and dependents, respectively. Furthermore, as discussed by Dryer (1992: 121-122), this challenges the notion of determiners as a unitary category for demonstratives and articles; as argued in the beginning of this section, distinct categories for articles and demonstratives are also supported by cross-linguistic evidence, whereas languages that mutually exclude articles and demonstratives in the same position, like half of the languages of my sample, are actually typologically rare.

As for the order of demonstratives with respect to numerals and nominal heads, Greenberg's Universal 20 states that:

> When any or all of the items (demonstrative, numeral, and descriptive adjective) precede the noun, they are always found in that order. If they follow, the order is either the same or its exact opposite. (Greenberg 1963: 68-69)

Using an undisclosed sample of languages, Cinque (2005) finds that only 14 out of the mathematically possible 24 orderings are actually attested and accounts for this in terms of movement from a universal underlying demonstrative < numeral < adjective < noun order; by contrast, Dryer (2018) claims that in a sample of 576 languages the attested orderings can be justified by describing the involved categories in semantic terms, rather than using syntactic categories as in Cinque's previous approach.

Owing to the confusion around the quantifier category, there are unsurprisingly very few studies on this category; Greenberg (1963) cautiously suggests that Universal 18 might be extended to non-numeral quantifiers, quoting Romance languages as an example.

With respect to the European languages object of this study, qualitative data on the orderings of the three categories can be collected from the World Atlas of Language Structure (WALS: Order of Demonstrative and Noun: Dryer 2013a) and two other works explicitly focusing on European languages (Dryer 2008 and Siewerska 1998); Table 1 reports this data, showing very little variability in the order of the articles, determiners and quantifiers. All languages with an article place it in the prenominal position, demonstratives are prenominal everywhere except for the Celtic languages and quantifiers are prenominal without exception. The only variability is represented by Romanian (ron; Indo-European, Italic) demonstratives, which Dryer classifies as 'mixed' according to a rule of thumb that states that "if the frequency of

the two orders is such that the more frequent order is less than twice as common as the other, the language is treated as lacking a dominant order for that pair of elements" (Dryer 2013b: 371).

As discussed elsewhere (Levshina et al. 2023; Talamo & Verkerk 2022), the type of data presented in Table 1, as well as the literature discussed above, suffers from what Wälchli (2009) addresses as 'data reduction'; continuous data, such as the frequencies invoked by Dryer in the quotation above, are reduced to categorical values. For instance, Table 1 uses three out of the six original values proposed by Dryer (2013a) for demonstrative-noun order: prenominal, postnominal and mixed; such an approach loses useful information, such as minor patterns that are not captured by methods like Dryer's rule of thumb or the quantity of variation behind a 'mixed' value.

Thanks to the availability of a growing body of computational resources like corpora and automatic parsers, the last decade has witnessed a number of quantitative studies using information-theoretic measures to capture word order variability (Futrell et al. 2015; Naranjo & Becker 2018; Alzetta et al. 2018; Gerdes et al. 2019; Levshina 2019).

However, these studies either do not consider any of the categories considered in the present study or conflate the three categories into a single category ('nominal modifier': Naranjo & Becker 2018: 94; 'determiner': Levshina 2019: 539). From a methodological perspective, these studies are problematic since (i) they do not provide a convincing match between cross-linguistically valid categories (comparative concepts: Haspelmath 2010; 2018) and instances of categories as found in corpora (tokens: Levshina 2019: 534) and (ii) they are based on non-comparable corpora (treebanks) which vary wildly regarding genre and size (Levshina et al. 2023: 32-34).

The first point stems from the fact that all studies, except for Levshina (2019), use only one type of annotation provided by the treebanks, namely, the syntactic relation between a dependent token and its head. As for the second point, treebanks are collections of manually or semi-automatically annotated texts, which are used to train Natural Language Processing tools, most notably, parsers; these linguistic resources are generally free from annotation errors, however their size is too small to incorporate semantic facts in the analysis. For this reason, Levshina (2021) uses a UD-parsed version of Leipzig Corpora to study, among other things, the relationship between the semantic properties of the verbal arguments and the order of subject and object.

Talamo & Verkerk (2022) introduce comparative concepts to study the order of four modifiers with respect to the nominal head; they show the implementation of these comparative concepts using two layers of annotation as provided by the UD framework, the syntactic relation layer and the Universal Parts-of-Speech layer, and manually-compiled list of lemmata, which are used to capture words from closed categories, such as articles, demonstratives and adpositions. Their approach allows to disentangle the category of determiners, showing, among other things, that the noun-demonstrative order is quite variable in two out of the 11 languages of their sample.

3. Data and Methodology

3.1 The CIEP+ corpus and the sample of languages

The corpus used in the present study is the Corpus of Indo-European Prose and More (henceforth: CIEP+), which has been developed from 2019 (Talamo & Verkerk 2022: 184-186). As the name suggests, the corpus currently features a collection of original versions and translations of 17 fiction books and 1 diary in 33 Indo-European languages, with a planned expansion to include translations from other linguistic families.

The criteria of selection of novels are quite simple: availability in a large number of languages and translations in a modern and accessible language variety. Both criteria are met by the so-called *best-seller* books, as their high demand means that are translated in several languages, using a variety that can be understood by the great majority of speakers. Talamo and Verkerk (2022) then included modern classics such as Marquez's *Cien Años de Soledad* (1967) and Eco's *Il nome della Rosa* (1980), as well as contemporary books such as the *Harry Potter* saga (1997-2007) and novels from Coelho, Musso and Süskind; in order to include minority languages, Talamo and Verkerk (2022) have selected less recent books such as Carroll's novels (*Alice's Adventures in Wonderland*: 1865; *Through the Looking-Glass*: 1871) and Saint-Exupery's *Le Petit Prince* (1943).

Since several translations are not (yet) available for all languages, I select for my sample 15 languages featuring the whole set of books (roughly 120,000 sentences or 2 million tokens for each language); these languages belong to the following branches:

- Germanic: Danish (dan; Indo-European, Germanic), Dutch (nld; Indo-European, Germanic), English (eng; Indo-European, Germanic), German (deu; Indo-European, Germanic);
- Hellenic: Greek (ell; Indo-European, Graeco-Phrygian);
- Romance: French (fra; Indo-European, Italic), Portuguese (por; Indo-European, Italic), Romanian (ron; Indo-European, Italic), Spanish (spa; Indo-European, Italic);
- Balto-Slavic: Bulgarian (bul; Indo-European, Balto-Slavic), Czech (ces; Indo-European, Balto-Slavic), Bosnian-Croatian-Serbian (henceforth: BCS; hbs; Indo-European, Balto-Slavic), Lithuanian (lit; Indo-European, Balto-Slavic), Polish (pol; Indo-European, Balto-Slavic), Russian (rus; Indo-European, Balto-Slavic).

The sample is completed by two minority languages belonging to the Celtic branch, Irish (gle; Indo-European, Celtic) and Welsh (cym; Indo-European, Celtic), each featuring five books (roughly 13,000 sentences, or 300,000 tokens).

The corpus is automatically parsed using Stanford Stanza[2] (Qi at al. 2020), which provides the traditional Natural Language Processing steps of sentence splitting, tokenization, lemmatization, as well as morphological and syntactic annotations using the Universal Dependency pre-trained models (de Marneffe et al. 2021).[3]

3.2 Determiners and quantifiers in European language: comparative concepts and the Universal Dependency framework

A challenge for typological studies is represented by the cross-linguistically valid definitions of the categories under scrutiny. These definitions, or 'comparative concepts', should rely on extra-linguistic factors, such as the semantics and the pragmatics of the categories, and should be different from language-specific categories, which are instead addressed as 'descriptive categories' (Haspelmath 2018; Croft 2016).

The usage of automatically annotated linguistic resources poses a series of additional problems, including the quality of the annotated data (Levshina et al. 2023: 29-32) and the cross-linguistic consistency of the annotation (Talamo & Verkerk 2022: 180-184).

[2] Version 1.3.0. https://stanfordnlp.github.io/stanza/

[3] Version 2.8. https://stanfordnlp.github.io/stanza/available_models.html

In what it follows, I exemplify both the theoretical and methodological matter on the categories of determiners and quantifiers, showing how comparative concepts can be implemented using the Universal Dependency (UD: de Marneffe et al. 2021) framework.

3.2.1 Comparative concepts

In this section, I discuss four comparative concepts and verify their adequacy for the 17 languages of my sample. Talamo and Verkerk (2022: Appendix C) propose the following two comparative concepts for the category of articles and demonstratives:

> Within a noun phrase, an ARTICLE[4] is a word that occupies a fixed position and expresses certain features of the nominal head, namely: (in)definiteness and/or specificity; additionally, an article may also signal deictic and/or anaphoric reference of the nominal head it modifies.
>
> Within a noun phrase, a DEMONSTRATIVE is a word that may vary its position and functionally characterizes the nominal head for deictic and/or anaphoric reference.

Central to the definition of ARTICLE is the notion of definiteness; in some languages and along the demonstrative-article grammaticalization path definiteness is found together with specificity (Himmelmann 2001: 831-832). Furthermore, deictic and anaphoric reference, which play a major role in the definition of DEMONSTRATIVE (Diessel & Coventry 2020: 1-2), are sometimes found "when articles encode meanings typically associated with demonstratives such as visibility or distance from a deictic center" (Himmelmann 2001: 837). Himmelmann attributes these 'deictic articles' to Salish and Wakashan languages, as well as to Austronesian languages (Himmelmann 2001: 837; see also Lyons 1999: 53-57); although the Indo-European languages of my sample lack 'deictic articles' i.e., dedicated markers for deixis and/or anaphora, their article systems are able to encode the opposition between 'familiar/unique reference' and 'non familiar/non-unique reference'.

[4] As a typographic and grammatical convention, I write comparative concepts using SMALL CAPS and treat them as singular nouns; language-specific categories such as English adjectives or Bulgarian demonstratives are written uncapitalized and are treated as plural nouns.

So far, I have discussed the two comparative concepts for their functions which, to a certain extent, tend to overlap; since the two comparative concepts are of the hybrid type (Haspelmath 2018: 86), they also include a formal aspect. It is precisely this formal aspect that distinguishes the two comparative concepts: an ARTICLE is a word occupying a fixed position, whereas a DEMONSTRATIVE is a word that may vary its position.

A number of languages from my sample do not fit, with different degrees, the ARTICLE comparative concept; in Balto-Slavic languages, (in)definiteness and specificity are coded by words belonging to other categories, such as adjectives or demonstratives (BCS: Alexander 2006: 20-21, Czech: Naughton 2005: 88; Lithuanian: Ramonienė et al. 2019: 49-52; Polish: Bielec 2012: 27; Russian: Timberlake 2004: 118-119), while in Bulgarian these features are coded by suffixes (Bulgarian: Antova & Boytchinova & Benatova 2002: 41-48). Celtic languages and Romanian meet the ARTICLE comparative concept only partially. In Irish and Welsh, a positive value of definiteness and/or specificity is coded by words occupying a fixed position, while indefinite nouns are bare nouns, cfr. Irish *an fear* 'the man' vs. *fear* 'man/a man' and Welsh *yr alarch* 'the swan' vs. *alarch* 'swan/a swan' (Stenson 2020: 183-185; King 2003: 28-30); in Romanian, fixed-positions words mark non-specific nouns and suffixes mark definite nouns, cfr. Romanian *un munte* 'a mountain' vs. *munte-le* 'the mountain' (Gönczöl-Davies 2008: 34-40).

The DEMONSTRATIVE comparative concept is valid for all languages of the sample, despite the different levels of deixis that a language may encode: (i) only one deictic value, as in French *ce* 'this/that' (Batchelor & Chebli-Saadi 2011: 609-612; see also Dryer 2007: 162-163, Diessel & Coventry 2020: 2-3); (ii) two deictic values, as in English *this* vs. *that*; (iii) three deictic values, as in Spanish *este* 'this' vs. *ese* 'that, close to the hearer' vs. *aquel* 'that, distant from both the speaker and the hearer' (Butt & Benjamins & Rodríguez 2019: 87-88).

Quantifiers can be analyzed cross-linguistically according to the following definition:

> Within a noun phrase, a QUANTIFIER is a word that may vary its position and functionally characterizes the nominal head for one of the following three types of non-numeral quantification: (i) distributive, (ii) proportional and (iii) amount-term.

The three types of quantification are described in Croft (2022: Glossary) and roughly correspond to the semantic classes discussed by Keenan (2012: 1-4). For the sake of convenience, I give here Croft's description of these three types:

- "distributive quantifier**:** a form that specifies the members of the set but treats them individually (that is, the predicate applies to the whole set by virtue of applying to the individual members of the set)". For instance, English *every* in *Every dog has fleas* indicates that each member of the *dog* set has *fleas*;
- "proportional quantifier: a form that specifies the set of instances as a proportion of the whole set of individuals/tokens of the type, or at least the contextually relevant whole set." For instance, English *few* in *few people were pleasantly surprised* indicates that a small proportion of the *people* set were pleasantly surprised;
- "amount-term quantifier: a form used to indicate an imprecise quantity for noncountable entities." For instance, English *some* in *pour me some wine* indicates an imprecise quantity of the mass noun *wine*.

Note that the first two types of quantification may be also expressed through numerals; these are excluded in the current study.

Although the consulted grammars use other terms to indicate the three types of quantification – only a reference grammar of Romanian (Dobrovie-Sorin & Giurgea 2013: 43-45) explicitly discusses proportional quantifiers – all sampled languages have words corresponding to the QUANTIFIER comparative concept.

For instance, the difference between distributive QUANTIFIER and proportional QUANTIFIER is described in Danish by Lundskær-Nielsen and Holmes (2010: 234) as an opposition between specific and universal application of the quantifier, which results in two different constructions.

(1) Danish (dan; Indo-European, Germanic; Lundskær-Nielsen and Holmes 2010: 234)

a. *Alle* *spillerne* *spillede* *dårligt.*
all players.DEF play.PST poorly
'All players played poorly.'

b. *Al* *magt* *til* *folket!*
all.M.SG power.M.SG to people
'All power to the people!'

In the example (1a), the *al* 'all' QUANTIFIER is followed by the definite form of *spillern* 'players', coding the distributive meaning – the action of playing poorly is predicated for each individual player; by contrast, in example (1b), the *al* 'all' QUANTIFIER agrees for gender and number with *magt* 'power' – the entire proportion of power should be given to the people.

Instances of amount-term QUANTIFIER are described in Danish by Lundskær-Nielsen and Holmes as "[they] can only modify non-count nouns to specify quantity or degree" (2010: 248), as in the following example using *lidt* 'some':

(2) Danish (dan; Indo-European, Danish; Lundskær-Nielsen and Holmes 2010: 248)

Må	*jeg*	*låne*	*lidt*	*sukker?*
May	I	borrow	some.N.SG	sugar.NCOUNT

'May I borrow some sugar?'

The amount-term QUANTIFIER applies to a non-countable entity – sugar – and the strategy employed by Danish is the lack of agreement between *lidt* and *sukker* 'sugar'.

The definition of NOMINAL HEAD involves two comparative concepts, one for the head and the other for the noun; the following comparative concept is based on Croft (definition of the head construction. This definition assumes that word categories are constructions of semantic classes (objects, actions, properties) and information packaging structures (propositional acts: reference, predication and modification; Croft 2001): "Within a noun phrase, a NOMINAL HEAD is the most contentful word that most closely denotes the function of referring as the phrase as a whole." (Croft 2022: Glossary).

This comparative concept encompasses all instances of HEAD governing a phrase that refers to objects i.e., a referring phrase; all languages of the sample have words corresponding to the definition of NOMINAL HEAD.

3.2.2 The UD framework and the List of Lemmata

The UD framework (de Marneffe et al. 2021) consists of several annotation layers spanning different levels of linguistic analysis; the annotation is performed at the token level and within sentence boundaries, with each token getting an incremental identification number (ID) starting from the first token of the sentence. For the purpose of the present study, I will employ two layers of UD annotation: (i) the Universal Parts of Speech (UPOS) layer, which annotates tokens for word categories

using a universal set of 17 tags[5] and (ii) the Relations (Rel) layer, which traces relations between tokens using their ID numbers and a list of 37 syntactic relations.[6] As the name suggests, syntactic relations are conceived of as dependencies, with a token acting as the head and another token acting as the dependent; furthermore, the structure of the annotation is hierarchical, with the sentence predicate acting as the main node (root). This is exemplified in Figure 1, which shows the analysis of the English sentence 'This corpus has several tokens.'. The main node of the sentence is the 'has' token and its immediate dependencies are the two arguments 'corpus' and 'tokens', which in turn are the head of 'this' and 'several', respectively. Figure 1 also shows the two layers of UD annotation, in which 'this' is annotated as a determiner (UPOS: DET) holding a determination dependency (Rel: det) with 'corpus' and 'several' is annotated as an adjective (UPOS: ADJ) holding an adjectival modification dependency (Rel: amod) with 'tokens'.

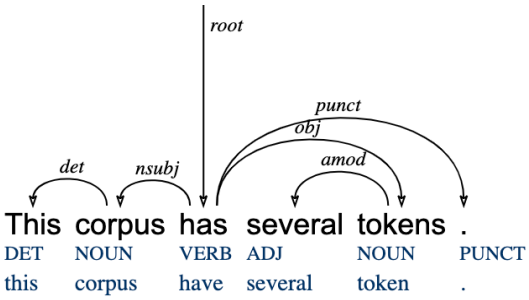

Figure 1: Analysis of the English sentence 'This corpus has several tokens' using the UD framework.

While the list of the UPOS tags is closed, the Rel list can be expanded using subtypes of existing relations; for instance, a number of languages uses a subtype of the determiner (det) relation in order to mark the relation between possessive pronouns and their head nouns, labelled 'det:poss'. Unfortunately, this has led to a proliferation of subtypes, which are quite often specific to a single language or a group of related languages.

Furthermore, the UD framework requires in principle a certain level of consistency between the UPOS and the Relation layer, with determiners (DET) performing determination (det), numerals (NUM) numeral modification (nummod), and so on.

[5] https://universaldependencies.org/u/pos/all.html

[6] https://universaldependencies.org/u/dep/index.html

Articles, demonstratives and quantifiers are treated as determiners (DET) in the UPOS layer and have a 'relation determiner' (det) "between a nominal head and its determiner".[7]

While the consistency between the UPOS and the Rel layer holds for manually-annotated treebanks, such as the ones available on the project website, it does not for corpora that are automatically parsed using parsers trained on these treebanks.
Beside an unavoidable rate of wrong annotations, casual inspection reveals several cases in which the determiner relation is associated with other UPOS tags rather than DET, most notably, adjectives (ADJ) and pronouns (PRON).

In order to reduce the effect of wrong and non-consistent annotations on the quality of data Talamo and Verkerk (2022) propose to add to the UPOS and the Relations annotation layers a third layer, the List of Lemmata (LoL) layer; this layer simply consists of a list of language-specific lemmata, which is compiled using reference grammars and consulting native speakers.[8] The LoL layer is then matched against the lemma annotation layer, which is also provided in the automatic annotation process.

C. Concept	UPOS	Relations	LoL
ARTICLE	DET	det det:predet	articles
DEMONSTRATIVE	DET PRON	det det:predet	demonstratives
QUANTIFIER	DET PRON (ADJ) (NOUN)	det det:predet det:numgov det:nummod (amod) (nmod)	quantifiers
NOMINAL HEAD	NOUN PROPN	-	-

Table 2: The comparative concepts and their implementation using the UD framework.

Table 1 shows the implementation of the four comparative concepts discussed in the previous section; this implementation is modular i.e., the three layers can be combined or excluded to obtain different results.

[7] https://universaldependencies.org/u/dep/det.html

[8] As pointed out by an anonymous reviewer, one may wonder to what extent the UPOS layer is still necessary after the introduction of the the LoL layer. To test this, I computed the entropy by combining the Rel and the LoL layers and keeping the UPOS layer only for the nominal heads; a paired *t-test* shows that the statistical difference between the mean entropy of this combination and of the Rel + UPOS + LoL combination for the three categories is not significant. The mean difference of entropy between the two combinations is .001 for the ARTICLE category, .002 for the DEMONSTRATIVE category and there is no difference for the QUANTIFIER category. As mentioned above, the UPOS layer is however still relevant to capture the nominal heads.

The UPOS tagset does not have specific tags for ARTICLE, DEMONSTRATIVE and QUANTIFIER; all these categories are conflated into the determiner (DET) tag, as described in the UD guidelines for the annotation of determiners;[9] additionally, I have included the PRON tag for DEMONSTRATIVE and QUANTIFIER, as adnominal forms are sometimes mistaken for pronouns by the parser. As for the NOMINAL HEAD, the category is implemented using the NOUN and PROPN tags.[10]

Along with the determiner relation, I have also included subtypes that are used in at least one language of the sample:

- det:predet, which is used in English to annotate the "relation between the head of an NP and a word that precedes and modifies the meaning of the NP determiner",[11] as in 'such a dangerous invention', where 'such' is a predeterminer for 'a';
- det:numgov and det:nummod, which are used in BCS, Czech and Polish to mark the difference between quantifiers that do not agree in number with their head (det:numgov) and quantifiers that do agree (det:nummod). For instance, contrast Czech *s několika složkami* 'with several components', in which *několika* 'several' does not agree for number with *složkami* 'components' and Czech *několik let* 'several years', in which *několik* agrees for number with *let* 'years'.

Finally, values given between brackets are used in combination with the LoL layer and only in the implementation of the QUANTIFIER comparative concept; these values include quite broad UPOS tags, adjectives (ADJ) and nouns (NOUN) together with the respective UD Relation, adjectival modification (amod) and nominal modification (nmod).

3.2.3 An information-theoretic approach to word-order

Following previous studies on word order (Montemurro & Zanette 2011; Koplenig et al. 2017; Levshina 2019; Talamo & Verkerk 2022), the amount of variability of instances of ARTICLE, DEMONSTRATIVE and QUANTIFIER is captured using information theoretic measures; more specifically, I employ Shannon's entropy, whose formula is given as follows:

[9] https://universaldependencies.org/u/pos/all.html#al-u-pos/DET

[10] https://universaldependencies.org/u/pos/all.html#al-u-pos/NOUN and https://universaldependencies.org/u/pos/all.html#al-u-pos/PROPN

[11] https://universaldependencies.org/en/dep/det-predet.html

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$$

where P represents the probability of a pattern of word order and n the possible number of patterns. Since we are concerned here with the order of the nominal head and one of its modifiers, n is set to 2.

The resulting entropy ranges from 0 i.e., only one of the two possible patterns is attested to 1 i.e., both patterns are attested with the same frequency.

For instance, there are 20 instances of prenominal demonstratives and 978 of postnominal demonstratives in the Greek translation of Marquez's *Cien Años de Soledad,* for a total number of 998 instances of DEMONSTRATIVE. The probability of the DEMONSTRATIVE-NOMINAL HEAD order is 0.02, while the probability of the NOMINAL HEAD-DEMONSTRATIVE ORDER is 0.98; the resulting entropy is obtained by the following equation:

$$H = -(0.02 \times \log_2 0.02 + 0.98 \times \log_2 0.98) = 0.141$$

4. Results

4.1. A quantitative overview

Figures 2 and 3 show the entropy of instances of ARTICLE, DEMONSTRATIVE and QUANTIFIER in the 17 languages, as captured by different combinations of annotation layers.

When the Relation layer is used alone, the three categories are indistinguishable from each other[12] and are conflated under the 'Determiners and Quantifiers' category, which is represented by the star shape in Figure 2; this is the methodological approach taken by most of the previous works using UD, as discussed in Sect. 2; this is also the approach capturing the highest level of entropy in all languages, with Bulgarian, Irish and Romanian exceeding the .5 value of entropy.

[12] Balto-Slavic languages are an exception here, as they use two Relation subtypes to annotate quantifiers. However, when taken together with the det Relation, the entropy of BCS, Bulgarian, Czech, Lithuanian, Polish and Russian quantifiers is very similar to the entropy of determiners.

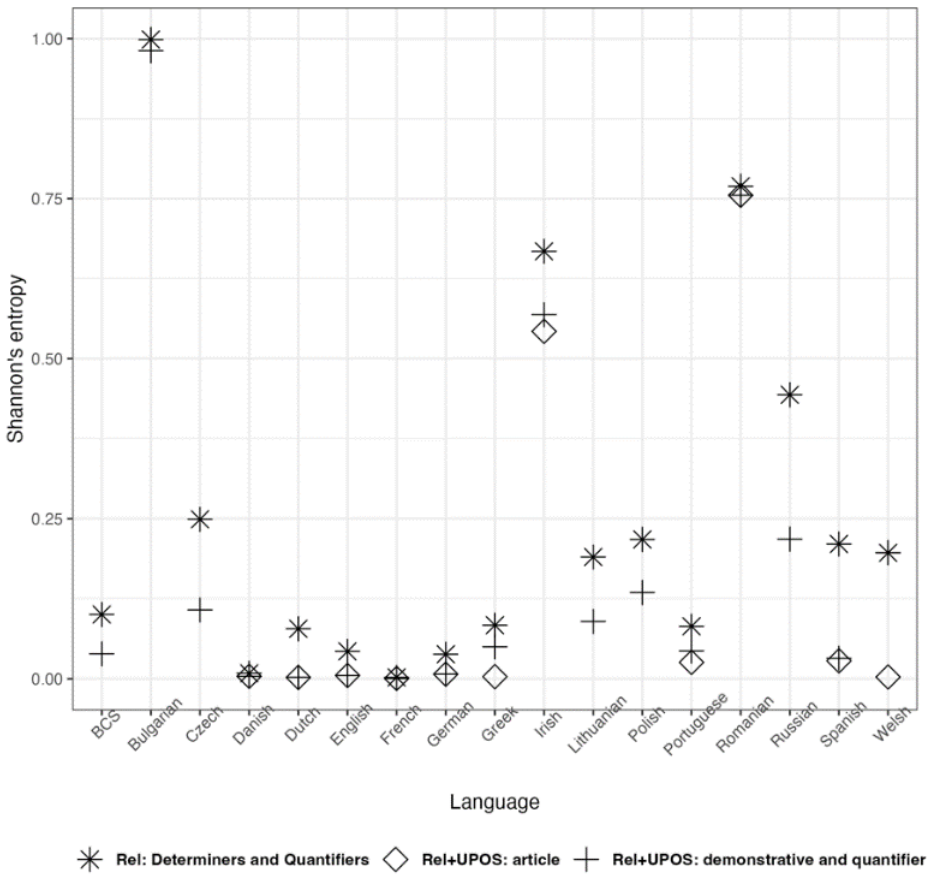

Figure 2: The entropy of 'Determiners and Quantifiers', as captured by the Relation layer only, the entropy of ARTICLE, as captured by the combination of the Relation and the UPOS layer and the entropy of 'Demonstratives and Quantifiers', as captured by the combination of the Relation and the UPOS layer.

The combination with the UPOS layer has the two-fold effect of separating the ARTICLE, which is identified by a diamond shape in Fig. 2, from the 'demonstratives and quantifiers' category, which is identified by a plus shape in Fig. 2, and reducing the entropy of all categories. This is particularly clear for languages already below the .5 value, which see their entropy reduced to quasi-null values.

The introduction of the LoL layer, which is combined with the other two layers in Figure 3, unpacks the 'demonstratives and quantifiers' category into the DEMONSTRATIVE and QUANTIFIER categories. The high entropy (.981) of the Bulgarian 'demonstratives and quantifiers' category is reduced to a quasi-null value (.01) for DEMONSTRATIVE and to .365 for QUANTIFIER, while the moderate entropy (.569) of the

Irish 'demonstrative and quantifiers' category raises to .848 for QUANTIFIER and drop to a quasi-null value for ARTICLE (.003).

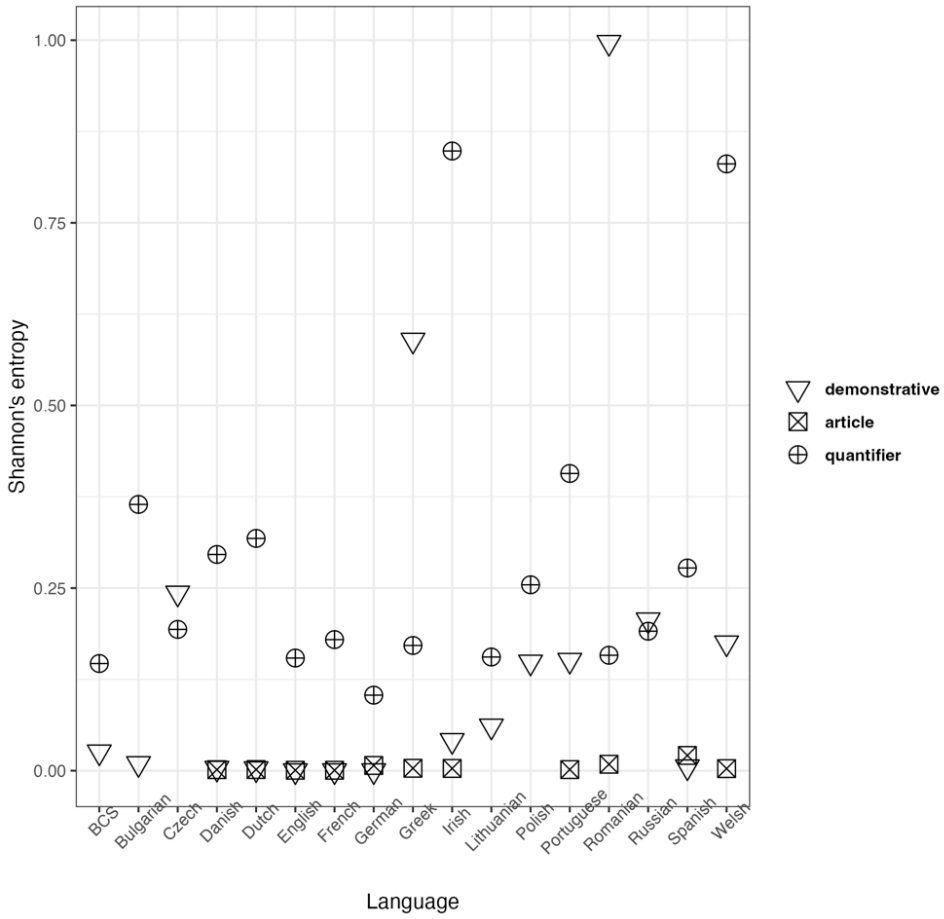

Figure 3 - The entropy of ARTICLE, DEMONSTRATIVE and QUANTIFIER, as captured by the combination of the Relations, UPOS and LoL layers.

In sum, there are four languages with entropy values above .500: DEMONSTRATIVE in Greek (.589) and in Romanian (.997), and QUANTIFIER in Irish (.848) and Welsh (.831).

As for Greek demonstratives, a slightly higher entropy is already observed by Talamo and Verkerk (2022) in the same corpus and is justified in terms of information structure. Demonstratives are generally prenominal in Greek, whereas postnominal demonstratives give an emphatic reading to the nominal head (Lascaratou 1998: 164), as in the following example from the Greek translation of Gabriel García Márquez *Cien años de soledad*, in which the rapid aging of Melquíades over a given period of time is emphasized:

(3) Greek (ell; Indo-European, Graeco-Phrygian) Gabriel García Márquez, *Cien años de soledad*, Greek trans. by Maria Palaialogou

Την	εποχή	εκείνη	ο	Μελκίαδες	γερνούσε	φανερά	από	τη
Tin	*epochí*	*ekeíni*	*o*	*Melkíades*	*gernoúse*	*fanerá*	*apó*	*ti*
ART.F	time.(F)	that.F	ART.M	Melquiades	age.IPFV.3SG	visibly	from	ART.F
μια	μέρα	στην	άλλη					
mia	*mera*	*stin*	*alli*					
one	day	to.the.F	the.F					

'At that time Melquíades was visibly aging from one day to the next.'

In the next section, I look more closely to the three other word order patterns with high entropy.

4.2. Some patterns of word order with high entropy

4.2.1 Variability of the DEMONSTRATIVE in Romanian: information structure or language register?

Instances of DEMONSTRATIVE in Romanian have the highest entropy (.997) across all languages and categories; out of a total frequency of 9086, 4248 demonstratives are prenominal and 4838 are postnominal, meaning that there is almost the same probability for both word order patterns.

According to Giurgea (2013: 160) pre-nominal and post-nominal positions are formally differentiated by the 'augmented' form[13] that demonstratives take in post-nominal position: *acest*-DEM *bărbat*-man 'this man' vs. *barbatul*-man.DEF *acesta*-DEM 'this man'; the high variability of Romanian demonstratives is evident in prose, where prenominal demonstratives "tend to be used with current discourse topics whereas postnominal demonstratives are preferred for rhematic and contrastive uses" (Giurgea 2013: 163). However, according to the same author, the position of the DEMONSTRATIVE as an information structure marker is lost in the modern-day speaking language and is replaced by an opposition of register: prenominal demonstratives are

[13] Dobrovie-Sorin and Giurgea (2013: 19) account for the difference between non-augmented and augmented forms in terms of phonological constraints.

used in the formal and literary variety, whereas postnominal demonstratives belong to the informal and colloquial Romanian.

Since CIEP+ is a corpus of literary texts, the high variability might be accounted for in terms of information structure; a look at the postnominal demonstratives in Romanian reveals that this strategy mostly codes a cohesion function, namely, anaphoric reference. This is illustrated by an example from *La Jeune Fille et la Nuit,* accompanied by the original sentence in French and the Greek translation; as mentioned above, Greek is the only other language of the sample showing a moderate entropy for the DEMONSTRATIVE, with a function similar to the one described for Romanian.

(4)

a. Romanian (ron; Indo-European, Italic) Guillaume Musso, *La Jeune Fille et la Nuit*, Romanian translation by Constantin Pistea

Știam	*foarte*	*bine*	*că*	*imaginea*	*aceasta*	*răspundea*	*aspirației*
know.PST.1SG	very	well	that	image.(F)	DEM.F	answer.PST.3SG	aspiration
mele	*acea*	*vreme.*					
My	DEM.F	time.(F)					

'I knew very well that this image answered my aspiration at that time.'

b. Greek (ell; Indo-European, Graeco-Phrygian) Guillaume Musso, *La Jeune Fille et la Nuit*, Greek translation by Maria Gourniezaki

Ήξερα	ότι	αυτή	η	εικόνα	ανταποκρινόταν	στις
Íxera	*óti*	*aftí*	*i*	*eikóna*	*antapokrinótan*	*stis*
know.PST.1SG	that	DEM.F	ART.F	image.(F)	answer.PST.3SG	to.the.F
προσδοκίες	εκείνης	της	εποχής.			
prosdokíes	*ekeínis*	*tis*	*epochís*			
expectation	DEM.F	ART.F	time.(F)			

'I knew that this image met the expectations of that time.'

c. French (fra; Indo-European, Italic) Guillaume Musso, *La Jeune Fille et la Nuit*, original French text

Je	*savais*	*très*	*bien*	*que*	*cette*	*image*	*répondait*	*à*
I	know.1SG.PST	very	well	that	DEM.F	image.(F)	anwer.3SG.PST	to

mon *aspiration* *d'* *alors.*
my aspiration of that.time
'I knew very well that this image answered my aspiration at the time.'

With respect to the second DEMONSTRATIVE, Romanian and Greek are aligned, in that they both translate with a prenominal distal demonstrative the French expression *d'alors* 'at that time'. By contrast, the French noun phrase *cette image* 'this image', which refers to a previously described image, is translated in Greek using the proximative demonstrative αυτή *aftí* 'this' in its unmarked (prenominal) position, while the Romanian translator uses the proximative demonstrative *acea* 'this' in postnominal position.

The high variability of the DEMONSTRATIVE might be also attributed to the large number of dialogues featured in several texts from CIEP+. Here, it is assumed that dialogues mimic, to a certain extent, the modern-day speaking language. If we look to the distribution of pre-nominal and post-nominal demonstratives across the texts of Romanian CIEP+ (Table 2) we have a partial confirmation of this hypothesis. For instance, the Harry Potter saga is aimed at a young audience, thus featuring a less formal language variety; the seven books from this saga contains 1536 prenominal demonstratives and 2429 postnominal demonstratives. A slight tendency toward a postnominal position of the demonstrative (231 pre-nominal vs. 279 post-nominal demonstratives) is also observed in the Romanian translation of Musso's *La jeunne Fille et la Nuit*, which belongs to a literary subgenre - the *novel noir* – that traditionally features a high amount of dialogues.

	Prenominal	**Postnominal**
Cien años de soledad	729	231
Adventures of Alice in Wonderland	56	74
Het Achterhuis	209	247
O Alquimista	125	103
La jeunne fille et la nuit	231	279
Il nome della rosa	721	801
Das Parfum	254	96
Le Petit Prince	30	50
Harry Potter and the Philosopher's Stone	159	179
Harry Potter and the Chamber of Secrets	142	166
Harry Potter and the Prisoner of Azkaban	122	285
Harry Potter and the Goblet of Fire	263	362

	Prenominal	Postnominal
Harry Potter and the Order of the Phoenix	356	574
Harry Potter and the Half-Blood Prince	298	438
Harry Potter and the Deathly Hallows	196	425
Through the Looking Glass	74	37
O Zahir	204	188
Βίος και Πολιτεία του Αλέξη Ζορμπά (Víos kai Politeía tou Aléxi Zorbá)	79	303

Table 2. The distribution of the position of Romanian demonstratives across the 18 books of CIEP+.

Finally, data from the largest UD treebank for Romanian (RoReRef: Barbu Mititelu et al. 2016) confirms that formal Romanian has a preference for the prenominal position for DEMONSTRATIVE; the entropy observed for DEMONSTRATIVE in this corpus, which features several genres such as law, medical, academic writing, is lower (.551), with 848 demonstratives in prenominal position and 124 in postnominal position.

4.2.2 Variability of the QUANTIFIER *in Celtic languages: artifacts or actual variation?*

The entropy of the QUANTIFIER is high for both Irish and Welsh; Irish has a value of .848, with 509 quantifiers in prenominal position and 1343 in postnominal position; Welsh a value of .831, albeit with fewer attested quantifiers i.e., 104 prenominal and 292 postnominal. In order to compare this data with other languages from the sample, it should be kept in mind that Irish and Welsh have only five books from the 18 featured in CIEP+, resulting in approximately one ninth of the total sentences, or one seventh of the total tokens. Furthermore, the performance of the parser for Irish and Welsh is lower with respect to the other languages of the sample;[14] accordingly, I additionally computed the frequency and the entropy of Irish and Welsh QUANTIFIER on the two UD treebanks available for these languages, UD Irish IDT and UD Welsh CCG, which are – at least partially – manually annotated. The entropy of quantifier in the two UD treebanks is higher than the entropy found for CIEP+: .97 for UD Irish IDT and .99 for UD Welsh CCG.

According to Stenson (2020: 188), the position of QUANTIFIER in Irish is lexically determined, as "most precede the noun in the same position as articles and pronominal possessors, but a few follow". Some quantifiers listed by Stenson are not

[14] See https://stanfordnlp.github.io/stanza/performance.html for a comparison between the performance of Stanza's pretrained models.

considered here, as they are either word combinations such as *go leor* ‘many, much, a lot’ and *ar fad* ‘all’, or are annotated by the parser as heads of nominal phrases, especially in prenominal position (see below).

	CIEP +		UD Irish IDT	
Lemma	**Prenominal**	**Postnominal**	**Prenominal**	**Postnominal**
beagán ‘a little’	0	15	0	0
céanna ‘same’	0	130	1	75
cuid ‘some, part of’	53	70	81	88
cúpla ‘a couple, a few’	5	5	1	2
éigin ‘some’	0	333	0	30
eile ‘other, another’	2	674	0	285
gach ‘every’	376	3	229	0
gach_uile ‘every’	2	0	17	0
mórán ‘many/much’	2	2	1	1
roinnt ‘some, a few’	1	1	4	5
tuilleadh ‘more’	0	6	0	1
uile ‘every’	68	104	16	44

Table 3. The distribution of Irish quantifiers at the lemma level and according to their position in CIEP + and in UD Irish IDT.

The distribution of the Irish quantifiers in CIEP + and in UD Irish IDT (Table 3) mostly reflects what Stenson (2020: 189-192) describes in her grammar, with a clear distinction between prenominal and postnominal quantifiers; an exception is represented by *beagán* ‘a little’ and *tuilleadh* ‘more’, which are described as prenominal quantifiers but appears only postnominally, and by some quantifiers appearing in both positions, most notably *cuid* ‘some, part of’ and *uile* ‘all’.

It seems, then, that a certain level of word order variability is also attested at the individual lemma level. However, a closer look to the token of these quantifiers reveals the fictious nature of this variation, with the possible exception of *cuid*.

Many instances of *beagán* and *tuilleadh* are not captured by the implementation of the QUANTIFIER comparative discussed in Sect. 3.2.2; when they appear in prenominal position, the two Irish quantifiers are annotated both in CIEP + and in the UD treebank as heads of nominal phrases; furthermore, the instances of postnominal quantifiers of *beagán* and *tuilleadh* are words modifying verbs or adjectives. Instances of *uile* in prenominal position are actually the two pronouns *uile dune* ‘everyone’ and *uile rud* ‘everything’, as well as other fixed expressions such as *uile cineál* ‘all kinds’

and *uile bhlas* 'all flavours'. According to Thurneysen (1990: 229), in Old Irish the position of *uile* is variable and the above-mentioned forms are allegedly relics of previous variability. Finally, *cuid*, along its usage as a prenominal quantifier, is also employed in possessive constructions, following pronominal possessors and preceding possessed objects, usually mass or plural nouns, e.g., *mo*.1SG *chuid airgid* 'my money'.

(5) Irish (gle; Indo-European, Celtic) J.K. Rowling, *Harry Potter and the Philosopher's Stone*, Irish trans. By Máire Nic Mhaoláin

Leag	*Mr*	*Ollivander*	*méar*	*fada*	*bhán*	*dá*	*chuid*	*ar*
laid	Mr	Ollivander	finger	long	white	3SG.POSS	CUID	on
an	*splanc*	*thintri*	*ar*	*éadan*	*Harry*			
the	flash	lightning	on	face	Harry			

'Mr. Ollivander laid his white long finger on the flash of lightning on Harry's face.'

The parser treats *cuid* as the postnominal modifier of the possessed object; for instance, in example (5) *cuid* is parsed as a nominal modifier (nmod) of *méar* 'finger'; this behavior is perhaps triggered by possessive constructions in which *cuid* is extended to non-pronominal possessors, but with a reversed word order, namely possessed object-*cuid*-possessor, as in example (6). This pattern may originate from a construction which "indicate(s) membership in a specific group" (Stenson 2020: 191) as in *Is inealtóir de chuid Aer Lingus é* 'He is an engineer from Aer Lingus'.[15]

(6) Irish (gle; Indo-European, Celtic) Saint-De-Exupery, *Le Petit Prince*, Irish trans. By Breandan O Doibhlin

Léiríodh	*dom*	*an*	*rún*	*eile*	*seo*	*de*	*chuid*	*an*	*phrionsa*	*bhig.*
Show.PASS	me	the	secret	other	DEM	of.it	CUID	the	prince	little

'I was shown this other secret of the Little Prince.'

As for Welsh, King (2003) describes the position of quantifiers as prenominal, with the *o* preposition preceding the noun in some cases e.g., *chwanag o de* 'some tea' but not in others: *sawl anifail* 'several animals' (125-126). Data from CIEP + and UD Welsh

[15] In an earlier draft of this paper, following Stenson (2020: 191), I have referred to *cuid* as a quantifier with partitive meaning; an anonymous reviewer suggests that its meaning might be better addressed as a part-whole relation, which is consistent with the group membership meaning discussed here.

CCG seem to contradict this statement, with more quantifiers in postnominal position than in prenominal position.

	CIEP +		UD Welsh CCG	
Lemma	**Prenominal**	**Postnominal**	**Prenominal**	**Postnominal**
digon 'enough'	11	34	0	1
gormod 'too much/many'	0	3	0	0
llawer 'a lot, much/many'	6	26	2	4
peth 'some'	28	169	2	8
rhagor 'more'	3	16	2	2
sawl 'several'	38	1	13	1
tipyn 'a (little) bit'	1	6	0	0
ychydig 'a (little) bit, a few'	17	37	5	3

Table 4. The distribution of Welsh quantifier at the lemma level and according to their position in CIEP + and in UD Welsh CCG.

However, these data should be handled carefully; the implementation of the QUANTIFIER category is at the same time too broad and too narrow. It is too broad as the nominal modification (nmod) relation captures several instances in which a word is not constructed as a quantifier; for instance, *peth* is used as the prenominal quantifier 'some' only colloquially (King 2003: 128-129), and is largely attested in Welsh CIEP + (169 occurrences) in postnominal position with its original meaning 'thing'; it is too narrow as, like in Irish, quantifiers are treated as heads of nominal phrases. Furthermore, the Welsh parser, probably because of its small training corpus, performs quite poorly, with several adjectives and/or verbs taken as nominal heads, an issue already encountered for some of the Irish quantifiers; Heinecke and Tyers (2019: 28-29) evaluate a parser trained on their treebank as "comparable with similar sized treebanks", however concluding that "the current 601 sentences may be a start, but do not cover enough examples to train a robust dependency parser". The current size of UD Welsh CCG does not also allow for meaningful comparison with the CIEP + data, as the frequency of the Welsh quantifiers is admittedly too low.

Differently from Irish, where there is sound evidence for lexically-based variation with some functionally and diachronically justified exceptions, data for Welsh quantifiers are either too noisy or too small to draw conclusions and the reported high entropy should, for now, be considered an artifact.

5. Conclusion

In the present paper I have analyzed the word order variation of articles, demonstratives and quantifiers in 17 European languages; these categories are notoriously hard to define cross-linguistically, and their variation has been poorly investigated in both qualitative and quantitative typological studies on word order.

Following previous quantitative studies, I treat word order variation as a continuous measure rather than a categorical one. However, with respect to previous studies, the methodology of the present paper aims to achieve a better match between typologically-adequate comparative concepts (category-like comparative concepts: Haspelmath 2018) and token-based comparative concepts, here represented by translations from the parallel Corpus of Indo-European Prose (CIEP). Following Talamo and Verkerk (2022), I combine the syntactic and part-of-speech layers of UD annotation with manually-crafted lists of lemmata in order to have a better representation of these categories at the token level.

The proposed methodology allows researchers to disentangle the entropy of the 'determiners and quantifiers' category, as captured by the single 'det' syntactic relations of the UD framework, into its three different components of ARTICLE, DEMONSTRATIVE and QUANTIFIER. Whereas the category of ARTICLE shows, as expected, no variation, DEMONSTRATIVE shows moderate-to-high values of entropy in Greek and Romanian, and the entropy of QUANTIFIER is high in Celtic languages; a closer look to these word order patterns reveals that the order of demonstratives in Romanian can be accounted for by principles of information structure, as previously shown by Talamo and Verkerk (2022) for Greek. The high entropy of Irish quantifiers is justified on lexical basis, while the high entropy of Welsh quantifiers turns out to be an artifact produced by the computational implementation of the QUANTIFIER category as well as by wrong annotations, which is due to the small training corpus available for Welsh.

The analysis of messy categories such as determiners and quantifiers is a testing ground for typological investigation using computational tools, such as the Stanza parser, UD models and parallel corpora; while these computational tools prove adequate for such a complex task in high-resource languages, the results for low-resource languages such as Welsh are not yet satisfactory enough. However, the development of new NLP tools and the extension of the UD framework to low-resource languages are rapidly evolving, and it will soon be possible to study (formerly) low-resource languages using quantitative typological methods such as the one discussed here.

Acknowledgements

Earlier versions of this work were presented at the 6th edition of the Using Corpora in Contrastive and Translation Studies conference (Bertinoro, September 2021) and at the 4th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (Seattle, July 2022). I thank the audience of both conferences for their comments. I am also indebted to Annemarie Verkerk, who has gone through the manuscript several times and provided precious help, as well as to two anonymous reviewers for improving this paper. All remaining errors are mine. This work is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project-ID 232722074 – SFB 1102.

To Dada and Giorgio, in loving memory.

Abbreviations

1 = 1sg person
3 = 3rd person
ART = article
ADJ = adjective
DEF = definite
DEM = demonstrative
DET = determiner
F = feminine
IMPF = imperfective
LOC = locative
M = masculine
N = neuter
NCOUNT = non countable
NUM = numerals
PART = partitive
PASS = passive
POSS = possessive
PL = plural
PRON = pronoun
PROPN = proper noun
PST = past
PUNCT = punctuation
SG = singular
UPOS = Universal Part of Speech

References

Alexander, Ronelle. 2006. *Bosnian, Croatian, Serbian, a Grammar*. Madison: The University of Wisconsin Press.

Alzetta, Chiara & Felice Dell'Orletta & Simonetta Montemagni & Giulia Venturi. 2018. Universal Dependencies and Quantitative Typological Trends. A Case Study on Word Order. In Nicoletta Calzolari et al., *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA). url: https://www.aclweb.org/anthology/L18-1719.

Antova, Evgenia & Ekaterina Boytchinova & Poly Benatova. 2002. *A short grammar of Bulgarian for English speaking learners* (2 ed.). Sofia: ABM Komers.

Arnaiz, Alfredo R. 1998. The main word order characteristics of Romance. In Siewerska, Anna (ed.), *Constituent Order in the Language of Europe*, 47-74. Berlin: Mouton de Gruyter.

Barbu Mititelu, Verginica & Elena Irimia & Cenel-Augusto Perez & Radu Ion & Radu Simionescu & Martin Popel. 2016. *UD Romanian RoRefTrees*. https://github.com/UniversalDependencies/UD_Romanian-RRT.

Batchelor, Ronald E. & Malliga Chebli-Saadi. 2011. *A Reference Grammar of French.* Cambridge: Cambridge University Press.

Bielec, Dana. 1998. *Polish: An Essential Grammar*. London & New York: Routledge.

Butt, John & Carmen Benjamin & Antonia Moreira Rodríguez. 2019. *A New Reference Grammar of Modern Spanish* (6 ed.). London & New York: Routledge.

Cinque, Guglielmo. 2005. Deriving Greenberg's Universal 20 and Its Exceptions. *Linguistic Inquiry* 36(3). 315–332.

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

Croft, William. 2016. Comparative concepts and language-specific categories: Theory and practice. *Linguistic Typology* 20(2). 377–393.

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

Diessel, Holger & Kenny R. Coventry 2020. Demonstratives in Spatial Language and Social Interaction: An Interdisciplinary Review. *Frontiers in psychology* 11.

Dobrovie-Sorin, Carmen & Ion Giurgea (eds.). 2013. *A Reference Grammar of Romanian: Volume 1: The Noun Phrase.* Amsterdam / Philadelphia: John Benjamins Publishing Company.

Dobrovie-Sorin, Carmen & Ion Giurgea. 2013. Introduction: Nominal features and nominal projections. In Carmen Dobrovie-Sorin & Ion Giurgea (eds.), *A Reference Grammar of Romanian: Volume 1: The Noun Phrase*, 1-48. Amsterdam / Philadelphia: John Benjamins Publishing Company.

Dryer, Matthew S. 1992. The Greenbergian Word Order Correlations. *Language* 68(1). 81-138.

Dryer, Matthew S. 1998. Aspects of Word Order in the Languages of Europe. In Anna Siewierska (ed.), *Constituent Order in the Languages of Europe*, 283-319. European Science Foundation Language Typology series. Berlin: Mouton de Gruyter.

Dryer, Matthew S. 2007. Lexical nominalization. In Timothy Shopen (ed.), *Language Typology and Syntactic Description. Grammatical Categories and the Lexicon (Second* Edition, 151–205. Cambridge: Cambridge University Press.

Dryer, Matthew S. 2009. The Branching Direction Theory of Word Order Correlations Revisited. In Sergio Scalise & Elisabetta Magni & Antonietta Bisetto (eds.), *Universals of Language Today*, 185-207. Berlin: Springer.

Dryer, Matthew S. 2013a. Order of Demonstrative and Noun. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. url: https://wals.info/chapter/88

Dryer, Matthew S. 2013b. Determining Dominant Word Order. In Matthew S. Dryer & Martin Haspelmath (eds.), *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. url: https://wals.info/chapter/s6

Dryer, Matthew S. 2018. The order of demonstrative, numeral, adjective, and noun. *Language* 94(4). 798-833.

Futrell, Richard & Kyle Mahowald & Edward Gibson. 2015. Quantifying Word Order Freedom in Dependency Corpora. In *Proceedings of the Third International Conference on Dependency Linguistics* (Depling 2015), 91–100. Uppsala, Sweden. Uppsala University, Uppsala, Sweden.

Gerdes, Kim & Sylvain Kahane & Xinying Chen. 2019. Rediscovering Greenberg's Word Order Universals in UD. In *Proceedings of the Third Workshop on Universal Dependencies* (UDW, SyntaxFest 2019). Paris, France: Association for Computational Linguistics, 124–131. doi: 10.18653/v1/W19-8015. url: https://www.aclweb.org/anthology/W198015.

Giurgea, Ion. 2013. The syntax of determiners and other functional categories. In Carmen Dobrovie-Sorin & Ion Giurgea (eds.), *A Reference Grammar of Romanian: Volume 1: The Noun Phrase,* 97-174. Amsterdam / Philadelphia: John Benjamins Publishing Company.

Gönczöl-Davies, Ramona. 2008. *Romanian: an essential grammar.* London & New York: Routledge.

Greenberg, Joseph H. 1963. Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements. In Joseph H. Greenberg (ed.), *Universals of Human Language,* 73-113. Cambridge, Mass: MIT Press.

Greenberg, Joseph H. 1978. Generalizations about Numeral Systems. In Joseph H Greenberg & Charles A. Ferguson & Edith A. Moravcsik (eds.), *Universals of Human Language*, Volume 3: Word Structure, 249–295. Stanford: Stanford University Press.

Haspelmath, Martin 2010. Comparative concepts and descriptive categories in cross-linguistic 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*, 83-114. Berlin: De Gruyter.

Hawkins, John A. 1983. *Word Order Universals*. New York: Academic Press.

Heinecke, Johannes & Francis M. Tyers. 2019. Development of a Universal Dependencies treebank for Welsh. In *Proceedings of the Celtic Language Technology Workshop*. Dublin: European Association for Machine Translation, 21-31. url: https://www.aclweb.org/anthology/W19-6904

Himmelmann, Nikolaus P. 2001. Articles. In Martin Haspelmath & Ekkehard König & Wulf Oesterreicher & Wolfgang Raible (eds.), *Language Typology and Language Universals* Vol. 1, 831-841. Berlin: Walter de Gruyter.

Holmberg, Andres & Jan Rijkhoff. 1998. Word order in the Germanic languages. In, Anna Siewerska (ed.), *Constituent Order in the Language of Europe*, 75-104. Berlin: Mouton de Gruyter.

Ioup, Georgette. 1975. Some universals for quantifier scope. In John Kimball (ed.), *Syntax and Semantics*, vol. 5, Academic Press, New York.

Keenan, Edward L. 2012. The Quantifier Questionnaire. In Edward Keenan & David Paperno (eds.), *Handbook of Quantifiers in Natural Language*. Studies in Linguistics and Philosophy, vol 90, 1-20. Dordrecht: Springer.

King, Gareth. 2003. *Modern Welsh: A Comprehensive Grammar*. London & New York: Routledge.

Koplenig, Alexander & Peter Paperno & Sascha Wolfer & Carolyn Müller-Spitzer. 2017. The statistical trade-off between word order and word structure - large-scale evidence for the principle of least effort. *PLoS ONE* 12(3)

Lyons, Christopher. 1999. *Definiteness*. Cambridge: Cambridge University Press.

Lascaratou, Chryssoula. 1998. Basic characteristics of Modern Greek word order. In Anna Siewerska (ed.), *Constituent Order in the Language of Europe*, 151-171. Berlin: Mouton de Gruyter.

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

Levshina, Natalia. 2021. Cross-linguistic trade-offs and causal relationships between cues to grammatical subject and object, and the problem of efficiency-related explanations. *Front. Psychol* 12.

Levshina, Natalia & Savithry Namboodiripad & Marc Allassonnière-Tang &
Mathew A. Kramer & Luigi Talamo & Annemarie Verkerk & Sasha Wilmoth & Gabriela Garrido Rodriguez & Timothy Gupton & Evan Kidd & Zoey Liu & Chiara Naccarato & Rachel Nordlinger & Anastasia Panova & Natalia Stoynova. 2023. Why we need a gradient approach to word order. *Linguistics* 61(4). 825-883.

Lundskær-Nielsen, Tom, & Philip Holmes. 2010. *Danish: A comprehensive grammar*. 2nd edn. Cambridge: Cambridge University Press.

de Marneffe, Marie-Catherine & Christopher D. Manning & Joakim Nivre & Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics* 47(2). 255-308.

Montemurro, Marcelo. A., & Damián H. Zanette 2011. Universal Entropy of Word Ordering Across Linguistic Families. PLoS ONE 6(5).

Naranjo, Matías Guzmán, & Laura Becker. 2018. Quantitative word order typology with UD. In Dag Haug & Stephan Oepen & Lilja Øvrelid & Marie Candito & Jan Hajič (eds.), *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)*, 91-104. Oslo: Linköping Electronic Conference Proceedings.

Naughton, James. 2005. *Czech: an essential grammar*. London & New York: Routledge.

Qi, Peng & Yuhao Zhang & Yuhui Zhang & Jason Bolton & Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In Dan Jurafsky & Joyce Chai & Natalie Schluter & Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations.* url: https://aclanthology.org/2020.acl-demos.14.pdf

Ramonienė, Meilutė & Joana Pribušauskaitė & Jogilė T. Ramonaitė & Loreta Vilkienė. 2019. *Lithuanian: A Comprehensive Grammar*. London & New York: Routledge.

Siewerska, Anna. (eds.). 1998. *Constituent Order in the Languages of Europe*. Berlin: Mouton de Gruyter.

Siewierska, Anna & Ludmila Uhlířová. 1998. Word order in the Slavic languages. In Anna Siewerska (ed.), *Constituent Order in the Language of Europe*, 105-149. Berlin: Mouton de Gruyter.

Stenson, Nancy. 2020. *Modern Irish: A Comprehensive Grammar*. London & New York: Routledge.

Talamo, Luigi & Annemarie Verkerk. 2022. A new methodology for an old problem: A corpus-based typology of adnominal word order in European languages. *Italian Journal of Linguistics* 34(1). 171-226.

Tallerman, Maggie. 1998. Word order in Celtic. In Anna Siewerska (ed.), *Constituent Order in the Language of Europe*, 21-45. Berlin: Mouton de Gruyter.

Thurneysen, Rudolf. 1990. *A Grammar of Old Irish, revised and enlarged edition, translated from the German by Daniel A. Binchy and Osborn Bergin*. Dublin: Dublin Institute for Advanced Studies.

Timberlake, Alan. 2004. *A reference grammar of Russian*. Cambridge: Cambridge: Cambridge University Press.

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

CONTACT

luigi.talamo@uni-saarland.de