DOI: 10.26346/1120-2726-197

A new methodology for an old problem: A corpus-based typology of adnominal word order in European languages

Luigi Talamo, Annemarie Verkerk

Language Science and Technology, Saarland University, Germany
<luigi.talamo@uni-saarland.de> <annemarie.verkerk@uni-saarland.de>

Linguistic typology is generally characterized by strong data reduction, stemming from the use of binary or categorical classifications. An example are the categories commonly used in describing word order: adjective-noun *vs* noun-adjective; genitive-noun *vs* noun-genitive; etc. Token-based typology is part of an answer towards more fine-grained and appropriate measurement in typology. We discuss an implementation of this methodology and provide a case-study involving adnominal word order in a sample of eleven European languages, using a parallel corpus automatically parsed with models from the Universal Dependencies project. By quantifying adnominal word order variability in terms of Shannon's entropy, we find that the placement of certain nominal modifiers in relation to their head noun is more variable than reported by typological databases, both within and across language genera. Whereas the low variability of placement of articles, adpositions and relative clauses is generally confirmed by our findings, the adnominal ordering of demonstratives and adjectives is more variable than previously reported.

KEYWORDS: adnominal word order, token-based typology, entropy, parallel corpus, Universal Dependencies, European languages.

1. Introduction

Classification and measurement in linguistic typology has mostly relied on discrete variables. Linguistic behavior is typically sorted into a fixed set of values, such as those for head- and dependent-marking of the direct object 'P': (a) P is head-marked; (b) P is dependent-marked; (c) P is double-marked; and (d) P has no marking (Nichols & Bickel 2013). The set of values can number four, such as in the example on marking in the clause, but it can feature three or more values, or it can be binary, when there are only two values, such as the binary opposition between having a distinction between 'and' and 'with', or not having such a distinction (Stassen 2013). These are just two examples, altogether, categorical variables are highly common in linguistic typology, as a cursory view of other chapters from the World Atlas of Linguistic Structure (henceforth WALS; Dryer & Haspelmath 2013) as well as other typological databases show.[1] Wälchli (2009) has demonstrated how this

(received May 2021)

common pattern of data reduction has resulted in typologists favoring the investigation of typological features that suffer least from being made categorical. Some typologies are naturally discrete. The example from Stassen (2013) on noun phrase conjunction seems to be clearly binary: either a language uses a nominal conjunction that also means 'with' or it does not. The same seems to be true for inventory-based phonological typology, i.e. either a language has a phoneme or it does not; features such as case, gender or number, i.e. either a language has case or it does not, and for certain aspects of grammar, i.e. a language either has symmetric negation, asymmetric negation, or both (Miestamo 2013).

But not all typologies are naturally discrete, as we detail in the following section. In this paper, we argue for a methodological approach to typology that is based on non-categorical measures; our proposal is part of a venture called 'token-based typology' (Levshina 2019, Levshina 2021). The name is inspired by Haspelmath (2018: 88), who distinguishes between typological methods that consider category-like concepts exemplified above and those that reflect concrete utterances or 'tokens' that emerge from choices in research design. Token-based typology can be done using experiments, in which visual stimuli form what Haspelmath (2018: 88) calls 'token-based comparative concepts' (for example, Slobin *et al.* 2011), or through translation, when comparison is enabled through the meaning of linguistic material elicited in a questionnaire or accessed through a parallel corpus that stays stable across translations (for example, Wälchli 2019).[2] Instead of using categorical variables, token-based typology measures cross-linguistic behavior using continuous measures. This emerges naturally as the unit of comparison are distributions of tokens (Levshina 2019: 534), which can be analyzed in a variety of ways, for instance by taking frequency (Bickel 2000), complexity (Hale 2016), entropy (Futrell, Mahowald & Gibson 2015), average surprisal (Hahn, Degen & Futrell 2021), or dependency length (Futrell, Levy & Gibson 2020). The main benefit of continuous, token-based measures is that they impose less harsh data reduction than categorical measures (Wälchli 2009), which implies that more data can be taken into account in a single analysis. Since the explanations of cross-linguistic distributions are many, and they intertwine, using token-based measures will ultimately result in a better assessment of why typological distributions are the way they are.

The present paper is organized as follows. In Section 2, we describe our implementation of a particular type of token-based typology, namely, corpus-based typology; in Section 3 we introduce a new datasource for token-based typology: CIEP+, the parallel Corpus of Indo-European Prose Plus. Our methodology and datasource can be in principle applied to several typological problems, but we have chosen to focus here on one of the oldest problems in typology, namely word order, for which we offer a case-study in Section 4. In this case-study, we compare word order variability using Shannon's entropy, an information-theoretic measure, in five modifier-noun pairs: adposition-noun, relative clause-noun, adjective-noun, article-noun and demonstrative-noun. In Section 5 we discuss pros and cons of our approach in the light of the case-study, while Section 6 concludes.

2. Corpus-based typology: A methodology at crossroads

2.1 Classification and measurement in typology

As exemplified in the Introduction, classification in typology has primarily been discrete, using binary or categorical classifications. There are three reasons for this; first, for some (perhaps one can even say many) phenomena, categorical classification simply fits most naturally (Wälchli 2009: 92), some examples of these were mentioned in the Introduction. Secondly, Cysouw (2005: 562) blames the preference for categorical classification on a need for simplification and "the widespread belief among linguists that everything essential in linguistic structure will be discrete". Thirdly, we must also relate this preference to the reluctance of many typologists to engage in statistical analysis beyond frequency counts, which is directly related to the fact that such analysis should deal with genealogical relations as well as areal convergence between languages, an issue that typologists have generally tried to solve through sampling. If a push for appropriate statistical testing in typology had been made prior to the 1990's, classification and measurement in typology would have been far more varied, as applying appropriate statistics forces one to take appropriate measurements (as noted by Cysouw 2005).

There are many examples where a categorical classification does not work, or where a continuous variable (see Cysouw 2005: 559 for examples) is a better descriptor of cross-linguistic variation. Some typologies are naturally gradient, such as Greenberg's morphological variable of synthesis (Greenberg 1960, see also Cysouw 2005: 559 and Levshina 2021: 5-6). The synthesis of a language is defined as the number of morphemes divided by the number of words in a text; low scores suggest lack of morphology, while high scores suggest great morphological complexity. Likewise, in some languages, auxiliary selection is dependent on the aspectual denotation of the verb (Sorace 2000) and should be measured as a gradient: a certain proportion of a language's verbal lexicon takes the auxiliary 'be', the rest takes the auxiliary 'have'. The third example, which is also the focus of our case-study, is word order. Word order has been treated from a predominant categorical perspective. In Dryer's influential typology of subject, object and verb (Dryer 2013g), languages which use primarily one word order (other word orders being either ungrammatical or only allowed in specific contexts) are rigid word order languages. Their dominant word order is classified into one of the six logical orders: SVO, SOV, VOS, OVS, OSV, and VOS. Hence, if a language in Dryer (2013g) is classified as 'SVO', that does not mean that no other word orders are attested. It means that 'SVO' is the dominant word order, which implies that it is either the only attested word order, or it is the more frequent one, such that the dominant word order occurs "twice as common as the next most frequent order" (Dryer 2013a). When this condition cannot be upheld, and all six logical orders are grammatical (at least in some context), the language may be said to have flexible word order. Dryer (2013g) does not make use of this value, rather, languages may be specified to have two dominant orders of subject, object and verb.

What do we miss out on if we apply categorical classification onto naturally continuous phenomena? This question is answered for word order by Levshina *et al. to appear*, rather than going over their arguments in detail, we want to make a point on word order that can easily be generalized. First of all, categorical classification of continuous phenomena results in data reduction, where the unknown factor is, "How much relevant information is lost?" (Wälchli 2009: 78). In Dryer's (2013g) typology of clausal word order described above, a language which reveals in a text count to have 40% SVO clauses, 20% SOV clauses, and 10% VSO, OVS, OSV, and VOS clauses, would be counted as having dominant SVO order.[3] Of course, this is an extreme example, nevertheless, categorization into so-called 'dominant' word order patterns hides potential variation in word order. Languages that display variable word order are not captured adequately by a categorical typology (see Wälchli 2009: 78-88 on object-verb order and see Appendix A in the Supplementary Material for concrete examples that will be explained further on, which show that for each of the five word orders studied in the current paper, there is at least one language which displays high word order variability).

As a concrete example we may take Dutch noun-genitive order. In Dryer (2013e), Dutch is stated to have noun-genitive order – the options in Dryer's (2013e) typology are (a) genitive noun order, (b) noun-genitive order, and (c) "both orders occur with neither order dominant". In fact, Dutch has four more or less commonly used genitive adnominal modifier constructions (Weerman & De Wit 1999), we include frequency counts in the Dutch part of the parallel corpus we introduce in Section 3:

I. *van*-genitive, as in *het boek van de leraar* 'the book of the teacher'; this is the most frequent construction with 27986 instances (90% of all genitives)
II. -*s*-genitive, as in *mijn leraars boek*[4] 'my teacher's book'; 2654 instances (9%)
III. possessive pronoun genitive, as in *de leraar z'n/zijn boek* 'the teacher's book'; the most uncommon pattern with 3 instances (0%)
IV. possessive *der* (archaic), as in *het boek der leraren* 'the book of teachers', less construed collocations are *rijk der natuur* 'realm of nature' or *Faculteit der Letteren* 'Faculty of Arts'; 441 instances (1%).

The first and last constructions, the *van*-genitive and possessive *der* indeed have noun-genitive order. But in the other two constructions, the -*s*-genitive and the possessive pronoun genitive, the head noun is preceded by the genitive. Similar to English's *of*-genitive and *'s*-genitive, the constructions are non-randomly distributed. For example, the prenominal genitive position is restricted to possessive pronouns, proper nouns, and other terms of address (i.e. *leraar* 'teacher', *buurvrouw* 'female neighbor', or *zusje* 'little sister'), which are interpreted as proper nouns (Weerman & De Wit 1999: 1167). A glance at the corpus queries mentioned above reveals that the -*s*-genitive is the most common possessive construction when the dependent is a proper noun, such as in *Harry's toverstok* 'Harry's wand'. The possessive pronoun genitive is rare in our corpus and may be more frequent in spoken and/or dialectal Dutch of the south. The classification in WALS hence misses out on at least one frequently-used construction (and its corresponding order), regardless of correctly construing noun-genitive order as the more frequent or 'dominant' order. Categorical classification that focuses solely on the most frequent patterns ignores constructions (with corresponding word orders) that are in all likelihood used by most speakers of Dutch on a daily basis. Classifying Dutch as a language in which "both orders occur" does not address the inherent problem with categorical classification, does not give us the usage ratio, and does not tell us anything about their non-random distribution regarding genitive type (proper nouns and other forms of address *vs* nouns and longer noun phrases).

The next question is of course what do we miss out on if word order variability or gradient phenomena in other domains of grammar are not taken into account. Here we must consider explanations in typology to bring home the relevance of the issue. Word order universals have been long explained in processing terms (Hawkins 1990, Hawkins 1994, Dryer 1992, Dryer 2009), however, a competing explanation for some word order universals is diachronic, rooted in grammaticalization processes (Bybee 1988; Collins 2019 and many others). An example by Dryer (2019) states that noun-adposition and verb-object orders correlate because verbs may grammaticalize into adpositions, and the order of the two components remains the same throughout the change. Dryer (2019) examines evidence for the grammaticalization account and finds that it fits some word order correlations, but not others. Using a token-based approach, Naranjo & Becker (2018: 100-101) find further indirect evidence for a diachronic explanation: they find a direct correlation between the proportion of object-verb/verb-object orders and the proportion of postpositions/prepositions in treebanks. If we ignore word order variability and label a language as having 'dominant' adposition-noun order, regardless of minor patterns of noun-adposition order, we cannot even begin to investigate the impact of 'processing' versus 'diachronic' explanations of typological distributions. Of course, a token-based approach alone cannot answer these questions either; we may need etymological analysis and/or historical reconstruction, for example, to identify the source of adpositions and genitives (Bybee 1988); and discourse analysis to explain clausal word order variability (Gundel 1988). Hence, while token-based typology cannot be the entire solution, it is definitely part of a solution for data reduction in typology and the consequences it has. All of these points have been made before. The current paper can be seen as an added voice to a chorus of increasing size, while making a specific contribution on the usage of cross-linguistic parsers by the Universal Dependencies project.

Rather than a single methodology, token-based typology can be described as a family of methods sharing three central features: taking observed tokens (instances) of particular linguistic structures as the basic data source, the already-mentioned treatment of such linguistic data as having non-discrete values, and the interdisciplinarity of the approach. Following Levshina (2021), we address the type of token-based typology described here as 'corpus-based typology'; in addition to typology, we understand corpus-based typology as a methodology involving at least three other disciplines, each fulfilling different purposes: corpus linguistics for data sources and methods, computational linguistics for the automatic annotation of corpora, and information theory for measures.

Given the necessity of tokens as a basic data-source, we may ask ourselves if token-based typology is a viable enterprise given that these materials, especially annotated corpora, are uncommon. Is it possible to apply token-based typology on a scale that is sufficient to make well-founded generalizations about human languages in general? First, we have to note that many researchers have already claimed that they have done so: the questionnaire-, experiment-, and corpus-based work in semantic typology and interaction spearheaded by the Language and Cognition group of the Max Planck Institute for Psycholinguistics can be construed as such (prominent examples are: Pederson *et al.* 1998, Majid, Enfield & Van Staden 2006, Majid, Boster & Bowerman 2008, Stivers *et al.* 2009, San Roque *et al.* 2015, Majid, Roberts *et al.* 2018, Seifart *et al.* 2018). Of course, samples in these studies are relatively small, generally not exceeding twenty genealogically and areally diverse languages. Larger samples have been researched using parallel corpora (Stolz & Gugeler 2000, Stolz, Stroh & Aina 2006, Wälchli & Cysouw 2012, Mayer & Cysouw 2014, Wälchli 2019). Additionally, non-parallel corpora, some with data on hundreds of different languages, do exist: the Leipzig Corpora Collection (Goldhahn, Eckart, Quasthoff *et al.* 2012); *Wikipedia Corpora*,[5] and *Multi-CAST* (Haig & Schnell 2021) or are being build (*100LC*: Bentz, Sozinova & Samardžić 2019). However, these massive corpora are not annotated with neither grammatical annotation nor with glosses, which makes them unavailable for research that requires such resources. In addition, methodology for parsing such corpora, for example zero-shot morphological and syntactical analysis (Basirat *et al.* 2019; Kann, Bowman & Cho 2020), are not available for use by the average typologist as they require considerable computational skills. Hence, while token-based typology definitely has been and can be used to make generalizations about human language, it has a long way to go in making resources and tools available and accessible. We hope to make a contribution through the new parsed parallel corpus presented here (see Section 3); in the remainder of this section we introduce information-theoretic measures and present computational tools for typological analyses, discussing their implementation as comparative concepts.

2.2 Information-theoretic measures

Information-theoretic measures have been used in order to study the cognitive and physiological load of language comprehension (see Hale 2016 for an overview), but they have also proven to be valid indexes of the complexity of the language structure (Bentz, Alikaniotis *et al.* 2017). The linguistic encoding of information has been studied both from a diachronic and synchronic perspective. By comparing different time slices of diachronic corpora using relative entropy (a quantitative measure of the ‘distance’ between two probability distributions), it is possible to model the evolution of linguistic features ranging from syntactic constructions (Degaetano-Ortlieb & Teich 2019) to the lexicon and discourse (Bochkarev, Solovyev & Wichmann 2014, Bizzoni *et al.* 2020). Synchronically, information content has been quantified across several languages at different levels of analysis (see Gibson *et al.* 2019 for a recent overview), including word distribution (Cohen Priva & Gleason 2016, Bentz, Alikaniotis *et al.* 2017), word length (Kalimeri *et al.* 2015), word and morph boundaries (Geertzen, Blevins & Milin 2016), inflectional morphology (Ackerman & Malouf 2013) and phonetic syllabic structure (Coupé *et al.* 2019). Additionally, word order has attracted much attention from an information theoretic perspective.

We focus here on studies of word order using entropy, as this is the measure we use in the current paper. Several studies (Montemurro & Zanette 2011, Koplenig *et al.* 2017, Levshina 2019) use Shannon’s entropy, which is defined as follows:

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

Formally, entropy measures the amount of information (which we can also think of in terms of uncertainty or surprise) involved in the value of a random variable or the outcome of a random process. In the current paper, as in the papers cited above, entropy is used as a measure of variability. We conceive of a word order pattern as having two possible representations, i.e. head-modifier and modifier-head; the upper bound of the summation, n, is set to 2, with P representing the probability of one of the two representations (x_i). If there is no variation, i.e. only one representation is attested, the entropy of the word order (X) is zero bit; on the extreme opposite, the maximum entropy value is one bit, which is reached whenever the two representations of word order have the same probability, i.e. 50% or $p = 0.5$.

We can distinguish two strands of studies on the entropy of word order. A first strand of studies use models based on individual words (Montemurro & Zanette 2011) or characters (Koplenig *et al.* 2017), without taking into account other information; these studies illuminate how the principles of information theory govern word order patterns, but say little on the formal and/or functional motivations of these patterns. A second strand of studies relies on layers of corpus annotation such as parts of speech or syntactic relations. For instance, Futrell, Mahowald

& Gibson (2015) use a conditional entropy measure on dependency corpora from HamleDT (Zeman, Dušek *et al.* 2014) and the Universal Dependencies project (Zeman, Nivre *et al.* 2020), claiming for a sample of 34 languages that variability in subject-object order is correlated with nominative-accusative marking.

The correlation between word entropy and case marking is further investigated by Levshina (2019), who considers 24 dependencies, again in the Universal Dependency framework, using Shannon's entropy as given above. Levshina (2019) claims that a significant correlation between word order entropy and syntheticity is found only in languages with VO order and rich morphological inflection, whereas OV languages, including those with rich morphological inflection, show less variability in word order. Moreover, Levshina (2019) explores the impact of lexical categories and functional motivations on word order; the lexical aspect is approximated by taking into account, both within and across languages, the entropies of individual modifiers and correlating those with the syntactic relations in which they appear. Functional motivations influencing word order variability, i.e. grammaticalization and processing constraints, are tested by individually correlating the entropy of verb and object (VO *vs* OV) with functional categories and modifier lengths. Both lexical and functional effects are confirmed by Levshina's (2019: 564) study; in particular, the following hierarchy of functional categories is supported by increasing levels of entropy:

> Function elements < Modifiers of a noun < Core Arguments < Adverbials = Obliques

where function elements such as adpositions and determiners have the lowest entropy, and adverbial/obliques such as pronominal phrases and adverbial clauses the highest.

As suggested by the above-discussed studies, information-theoretic measures such as entropy are taking their place next to categorical and continuous measurement in typology. Entropy has an important role to play, as recent work has uncovered that a single measurement on word order variability is highly informative with regard to its relation to inflectional morphology (see above, Levshina 2019), grammaticalization (see above, Naranjo & Becker 2018), and possibly information status (ongoing work by authors).[6] Rather than modeling one or several token-based ratios, entropy allows for a single measure of variability with inherent value. Its usage extends beyond word order variability, i.e. variability in overt arguments, presence of case marking, realization of phonemic variants can all be expressed using entropy. Furthermore, as mentioned at the beginning of this section, information theoretic measures have been used to quantify on-line processing and parsing which, in turn, represent some of the functional explanations that typologists have been putting forth for decades.

Both frequency-based measurement and entropy can help us quantify word order variability better than using categorical classifications. As discussed above, Dryer (2013g) deals with word order variability using the following rule of thumb:

> The rule of thumb employed is that if text counts reveal one order of a pair of elements to be more than twice as common as the other order, then that order is considered dominant, while 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 2013g)

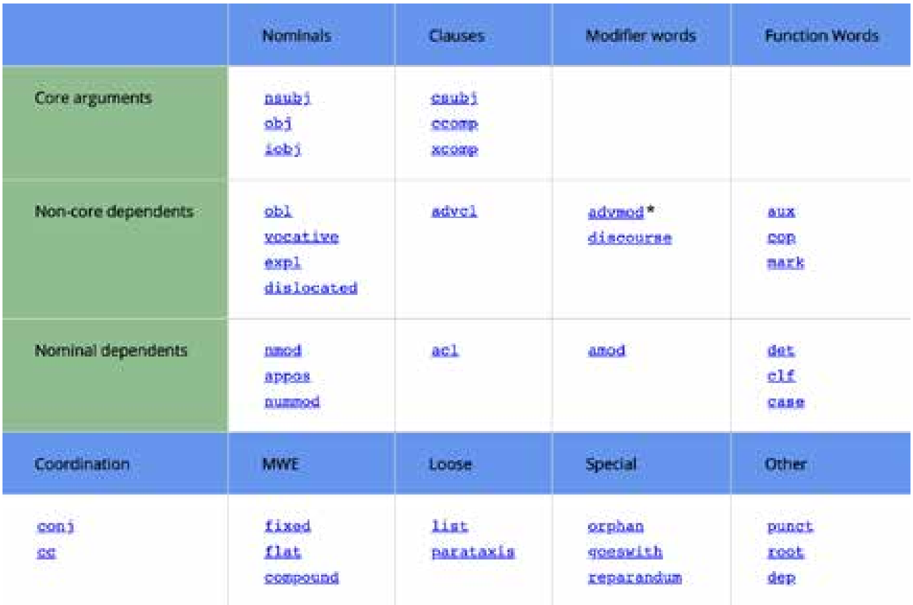
Dryer's rule of thumb is interpreted differently depending on the number of values included in the categorical variables: as we already saw, for Dryer's (2013g) typology of clausal word order described above, a language which reveals in a text count to have 40% SVO clauses where all other orders appear 20% of the time or are less frequent will be classified as having dominant SVO order. In this paper, we concern ourselves exclusively with binary variables. In this case, Dryer's rule of thumb establishes a dominant word order if it is represented by a frequency of 67% or more; this corresponds to a Shannon's entropy value of 0.915, which in fact depicts a very high level of variability and a far cry from the 'dominance' of a word order over the other. In this paper and in general, we would posit that the presence of a 'minority' word order pattern with a frequency over 5% is worth describing and bringing into an analysis. However, this value may be too high for diachronic analysis, for example in analyzing if (the direction of) word order change can be related to word order variability. It may be highly appropriate to include patterns that appear less frequently than 5% of the time, especially if the data is not noisy. The main argument that we want to argue for is to include 'minority' patterns in word order and other variables in appropriate ways.

2.3 Devising computational tools for linguistic typology

Traditional computational tools for the automatic processing of languages (also known as Natural Language Processing/NLP tools), such as parts-of-speech taggers (POS taggers) and dependency parsers have been developed for single languages, using the computational equivalent of Haspelmath's (2018) descriptive categories. The Universal Dependencies project (henceforth UD; Marneffe, Manning *et al.* 2021), which has been developing since 2014, is one of the few attempts to elaborate a framework that is suitable for cross-linguistic investigations.[7] Despite its name, the UD project does not only cover syntactic relations ('Universal Dependency Relations': UD Relations), but includes the 'Universal POS tags', which describe lexical categories and the 'Universal features', attempting to provide an exhaustive list of every morphosyntactic category found in human language (Marneffe, Manning *et al.* 2021: 260-265).

The UD project is built on previous projects: the Universal Stanford Dependencies are the basis for UD Relations (Marneffe, Dozat *et al.* 2014), the Universal POS tags extends Google Universal Parts-of-Speech Tagset (Petrov, Das & McDonald 2012) and the Universal features draw on the Interset universal set (Zeman 2008). We start by discussing UD Relations, the main annotation layer used in the present work.

UD Relations are a taxonomy of syntactic relations between words (Marneffe, Manning *et al.* 2021: 259-260, 265-268). The taxonomy is organized into two layers: a closed core of 37 universal relations and an open layer of relation subtypes. The core is designed as a matrix, as depicted in Figure 1. In this matrix, rows represent functional categories with respect to the head: core arguments of clausal predicates, non-core dependents of clausal predicates and nominal dependents, while columns describe structural categories of the dependent: nominals, clauses, modifier words and function words. UD Relations are found at the intersection of functional and structural categories, with the exception of a dozen of relations that are described only according to their structure, as they "are not dependency relations in the narrow sense",[8] for instance: multi-word expressions (called 'fixed'), coordinated items ('coord'), punctuation ('punct') and the head of the sentence, the root element ('root'). The additional layer features over 300 relation subtypes,[9] which are extensions of the UD Relations allowing "further distinctions or to capture language-specific constructions" (Marneffe, Manning *et al.* 2021: 265). As opposed to the core, this layer is open to new relation subtypes, meaning that these subtypes are specific to a subset of languages and, sometimes, to just one language.



	Nominals	Clauses	Modifier words	Function Words
Core arguments	nsubj obj iobj	csubj ccomp xcomp		
Non-core dependents	obl vocative expl dislocated	advcl	advmod* discourse	aux cop mark
Nominal dependents	nmod appos nummod	acl	amod	det clf case
Coordination	MWE	Loose	Special	Other
conj cc	fixed flat compound	list parataxis	orphan goeswith reparandum	punct root dep

Figure 1. Universal Dependency Relations, taken from <universaldependencies.org/u/dep/index.html>.

An approach based on UD Relations and, more in general, on Universal Dependencies, seems to be robust enough to study not only word order, but a greater variety of linguistic phenomena such as case markers (Levshina 2019 and see below) and analytic (free-standing) function words (Levshina 2021); furthermore, UD-based computational tools have been used to assess the genealogical similarities between languages (Chen & Gerdes 2018) and to quantify different measures of cross-linguistic complexity (Berdicevskis *et al.* 2018). However, we argue here that UD-based computational tools for typological analysis need some refinements. We are mostly concerned here with the source of data and with the quality of the analysis. Most of the previous studies addressing word order are based on UD treebanks, i.e. collections of linguistic data that are, at least in part, manually annotated or are carefully revised after automatic annotations (Futrell, Mahowald & Gibson 2015, Naranjo & Becker 2018, Alzetta *et al.* 2018, Gerdes, Kahane & Chen 2019). This methodology produces highly reliable results, as the analyzed data is generally very clean (Alzetta *et al.* 2018); on the other side, as treebanks highly differ across languages for size and text genre, UD treebanks are not readily comparable. In this paper we try to address the 'comparability' issue by using a collection of parallel texts that are automatically annotated by parsers trained on the UD treebanks. However, as the quality of parsing varies from language to language, we are faced with the problem of filtering out a certain amount of erroneous annotation, or 'noise'. A way to reduce the noise is by introducing additional annotation layers; in the case of UD Relations (syntactic relations), we can do this by additionally combining them with POS tags.

The Universal Dependencies POS tagset, the UPOS, provides a set of 17 lexical categories; it aims to be universal and should represent the morpho-syntactic, token-level counterpart to the UD syntactic Relations. This implies that in principle, some UD Relations should match corresponding UPOSes; for instance, the 'case'[10] relation should match only with modifiers with the adposition UPOS tag, ADP, while the 'amod' relation should match only with modifiers with the 'ADJ' UPOS tag. This is generally true for UD treebanks, but unfortunately UPOS annotation is not free of mistakes either. Filtering out erroneously assigned modifiers on the UD Relation level by using UPOS may help to reduce to a certain amount the annotation errors; but at the same time we risk to exclude correct modifier analysis due to incorrect UPOS assignment.

Furthermore, whereas UD Relations are quite consistent across languages, mostly covering the same language-specific syntactic relations, the 'universality' of a parts-of-speech tagset (Marneffe, Manning *et al.* 2021: 261) is much harder to achieve, as word categories, especially minor word categories such as articles or quantifiers, are language-specific (see Croft *et al.* 2017 for reworking UD parts of speech using information packaging). And it is exactly in minor word categories that we find the highest number of inconsistencies; for instance, proform categories, such as interrogative, personal, possessive and relative pronouns are tagged either as DET (Determiner) or PRON (Pronoun) by UPOS sets for different languages. We propose to address this by restricting the word categories we take into account; however, even when using more narrow comparative concepts, we sometimes find that the UPOS tagset is still not cross-linguistically consistent. To take just an example from the sample of languages featured in our case-study (Section 4), we have restricted the set of the determiners to the subset of article and demonstrative, which are tagged in all languages but one by the DET UPOS tag; the exception is represented by Welsh, in which the DET tag only covers the article category, while demonstratives are annotated with the PRON UPOS tag.

In order to investigate closed word categories such as adposition, article, demonstrative, as well as different kinds of pronouns and quantifiers, we propose to use language-specific lists of lemmata, which are compiled combining information from corpus-based analyses and reference grammars, thus integrating classical grammar-based typology into the framework of token-based typology, and, where possible, exploit a language-specific annotation layer of Universal Dependencies, the lemma field.[11] This multi-layered approach for the analysis of corpus-based data, consisting of different combinations between the UD Relations, UPOS tags and language specific lists of lemmata, is tested in Section 4. By doing this, this paper departs from previous studies in token-based typology (e.g. Futrell, Mahowald & Gibson 2015; Levshina 2019), which use different data and do not consider combining layers of UD annotation in this way.

2.2 CIEP+: the parallel Corpus of Indo-European Prose Plus

We introduce the parallel Corpus of Indo-European Prose Plus (henceforth CIEP+, /kiːp plʌs/), a project currently in development. We aim to include 43 languages in CIEP+, a balanced sample of 33 Indo-European languages, as well as 10 non-Indo-European languages, compiling translations of 18 literary works. More information on CIEP+ can be found in the Supplementary Material (Appendix A) and on the authors' website.[12]

As the corpus name suggests, all the texts are of a prosaic nature, belonging to the fiction and epistolary genres. In some cases, but not all, literary texts can be close to spoken and vernacular lects, showing features such as (direct) dialogues, informal lexicon and second-person addressees (similar to film subtitles, Levshina 2017). CIEP+ includes books from the *Harry Potter* series, which include many dialogues, as well as Garcia Márquez's *Cien años de soledad,* which includes hardly any dialogue. Existing parallel corpora tend to be biased "toward religious or legalese registers" (Wälchli 2007: 132, see for example the Parallel Bible Corpus, Mayer & Cysouw 2014 and EuroParl, Koehn 2005), sourced from the web (Goldhahn, Eckart, Quasthoff *et al.* 2012), or also of the literary genre (Stolz & Gugeler 2000; Stolz, Stroh & Aina 2006; Waldenfels 2006). With CIEP+, we build on these in terms of size (in comparison with the work of Stolz and colleagues) and number of languages (in comparison with von Waldenfels' corpus PARASOL).

We aim to compile a corpus of contemporary language use and have striven to include original texts and translations from the 1940s onward and possibly even later (min.: 1865, mean: 1995, max.: 2019). The outliers here are Carroll's novels, *Alice's Adventures in Wonderland* and *Through the Looking-Glass and What Alice Found There,* whose original texts date back to 1865 and 1871, respectively, and Musso's novel, *La jeune fille et la nui*t, which was published in its original French in 2018; the translations of Carroll's novels are however quite recent (min.: 1914, mean: 1984, max.: 2010), and *La jeune fille et la nuit* was translated almost immediately (2018-2019) into the other languages of our sample, with the exception of Portuguese and Welsh, for which there are no translations. CIEP+ is currently biased towards translations from English, as half of the texts are originally written in this language, with the remaining half divided between seven additional original languages (Dutch, French, German, Greek, Italian, Portuguese, and Spanish).

In this paper, we use a preliminary version of CIEP+, the Indo-European part of the corpus, which we also call CIEP. We include 18 different novels in 11 modern Indo-European languages from five subgroups: Celtic (Welsh), Germanic (Danish, Dutch, German, English), Hellenic (Modern Greek), Romance (French, Italian, Portuguese, Spanish), and Slavic (Polish), for an approximate number of 120,000 of sentences and 2M of tokens for each language; nearly all the novels (17 out of 18; cf. Appendix A in the Supplementary Material) are translated into each one of these languages, with the exception of Welsh, which features 5 different novels, as the remaining novels are not translated into Welsh (Welsh CIEP: 342,074 tokens).

CIEP+ is a work in progress. The corpus is currently being built, Germanic and Romance languages are currently over-represented and, among the Indo-European languages spoken in Europe, Baltic languages as well as Albanian, are not represented at all; finally, a corpus featuring prominently Indo-European languages should feature languages from the Indo-Iranian subgroup, alone accounting for almost two thirds of the whole family. In order to solve these issues, we ultimately aim for a phylogenetically balanced sample of thirty-three modern Indo-European languages, representing all nine living subgroups (Celtic, Romance, Germanic, Balto-Slavic, Indo-Aryan, Iranian, Greek, Armenian, and Albanian, see Clackson 2007 and Appendix A in the Supplementary Material). The ten non-Indo-European languages included are sampled to incorporate genealogical and areal diversity. Samples of this kind have been used in typology before (Futrell, Levy & Gibson 2020; Futrell, Mahowald & Gibson 2015; Hahn, Degen & Futrell 2021; Levshina 2019, 2021; Stolz & Gugeler 2000), but of course all involved in this work know that these are highly biased samples, with a huge bias towards the European Indo-European languages; languages from the Americas typically do not feature in this work. A more appropriate sample is possible for some of the texts included in CIEP+: the first *Harry Potter* book has been translated in over 75 languages; *Alice's Adventures in Wonderland* in over 200 languages; and *Le Petit Prince* in over 450 languages. We hope that CIEP+ is the start of a parallel corpus with a bigger and more varied language sample. We plan to use it in various forms: as a parallel sentence-aligned corpus with traditional tokenization, lemmatization and POS-tags, as a parallel corpus parsed according to the UD specifications, as a starting point for computational models, and so on.

3. Case-study: A corpus-based typology of adnominal word order in European languages

3.1 Data and Methods

We present in this section a case-study on adnominal word order, considering a preliminary version of CIEP. The corpus was parsed using UDPipe 1 (Straka & Straková 2017) with UD 2.5 models (Zeman, Nivre *et al.* 2020) specific to each of the languages of the sample (see Appendix C in the Supplementary Material for the list of models[13]). The UD-parsed parallel corpus is analyzed using the *pyconll* library.[14] Following the previous case-studies in Levshina (2019), we apply Shannon's entropy as a measure (a) for language-internal variability; (b) for comparing languages in terms of variability and (c) for comparing word orders, i.e. assessing which word order allows for greater or smaller variability across the board.

In order to illustrate our procedure, we take a long and descriptive sentence from Eco's *Il nome della rosa*, one of the literary works included in CIEP, and we calculate the entropy of adjective and noun order in the English and French translations:

# sent_id = NomeRosa_English_s3754

# text = As a little drop$_{p(1\cdot00)}$ of water added to a quantity of wine is completely dispersed and takes on the color and taste of wine, as red-hot iron$_{p(1\cdot00)}$ becomes like molten fire$_{p(1\cdot00)}$ losing its original form$_{p(1\cdot00)}$, as air when it is inundated with the sun's light is transformed into total splendor$_{p(1\cdot00)}$ and clarity so that it no longer seems illuminated but, rather, seems to be light itself, so I felt myself die of tender liquefaction$_{p(1\cdot00)}$, and I had only the strength left to murmur the words of the psalm: "Behold my bosom is like new wine$_{p(1\cdot00)}$, sealed, which bursts new vessels$_{p(1\cdot00)}$," and suddenly I saw a brilliant light$_{p(1\cdot00)}$ and in it a saffron-colored form$_{p(1\cdot00)}$ which flamed up in a sweet$\varnothing_{p(1\cdot00)}$ and shiningfire$_{p(1\cdot00)}$, and that splendid light$_{p(1\cdot00)}$ spread through all the shining fire$_{p(1\cdot00)}$, and this shining fire$_{p(1\cdot00)}$ through that golden form$p_{(1\cdot00)}$ and that brilliant light$p_{(1\cdot00)}$ and that shining fire$_{p(1\cdot00)}$ through the whole form$_{p(1\cdot00)}$.

> \# sent_id = NomeRosa_French_s4436
> \# text = Comme une petite goutte$_{p(0\cdot3)}$ d'eau instillée dans une grande quantité$_{p(0\cdot3)}$ de vin se dissipe tout à fait pour prendre couleur et saveur de vin, comme le fer incandescent$_{p(0\cdot7)}$ et $\emptyset$ enflammé$_{p(0\cdot7)}$ devient tout semblable au feu, perdant sa forme primitive$_{p(0\cdot7)}$, comme l'air inondé par la lumière du soleil est transformé en la plus grande splendeur$_{p(0\cdot3)}$ et en la même clarté$_{p(0\cdot3)}$, au point de ne pas paraître illuminé mais être lumière lui-même, ainsi je me sentais mourir de tendre liquéfaction$_{p(0\cdot3)}$, si bien qu'il ne me resta plus que la force de murmurer les paroles du psaume : " Voici : ma poitrine est comme le vin nouveau$_{p(0\cdot7)}$, sans ouverture, qui brise les outres neuves$_{p(0\cdot7)}$ " , et aussitôt je vis une éclatante lumière$_{p(0\cdot3)}$ et en elle une forme couleur du saphir qui s'enflammait tout entière d'un feu rutilant$_{p(0\cdot7)}$ et très $\emptyset$ suave$_{p(0\cdot7)}$, et cette lumière splendide$_{p(0\cdot7)}$ se dissémina complètement dans le feu rutilant$p_{(0\cdot7)}$, et ce feu rutilant$_{p(0\cdot7)}$ dans cette forme resplendissante$_{p(0\cdot7)}$ et cette lumière éclatante$_{p(0\cdot7)}$ et ce feu rutilant$_{p(0\cdot7)}$ dans la forme tout entière$_{p(0\cdot7)}$.

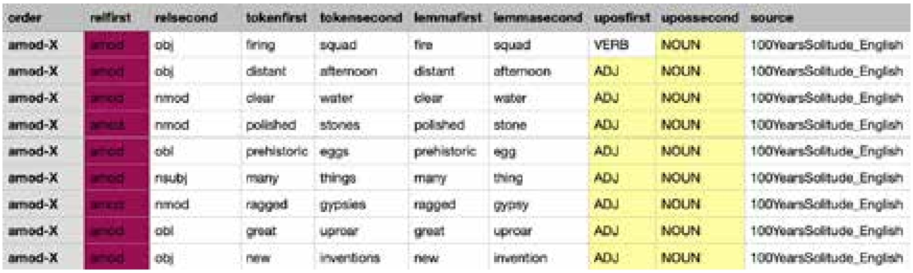
The English sentence contains 19 instances of the adjective-noun order, whose pattern is always modifier-head; accordingly, we assign to all instances a probability of 1.0 (or 100%), with a resulting entropy $H = -(1 \times \log_2(1) + 1 \times \log_2(1)) = 0$ i.e. there is no variation. As for the French sentence, we find 20 instances of the adjective-noun order, distributed across 6 instances with the modifier-head pattern (30%: $p(0.3)$) and 14 instances with the head-modifier pattern (70%: $p(0.7)$); the entropy is given by $H = -(0.3\times \log_2(0.3) + 0.7\times \log_2(0.7)) = 0.88$. These calculations are done on the entire corpus for a set of five adnominal word orders: adposition-noun, relative clause-noun, adjective-noun, article-noun and demonstrative-noun. Several other adnominal modifiers are excluded from the present study: classifiers, as the languages of the sample do not feature classifiers; pronouns and quantifiers, the reasons for which are explained in Appendix D in the Supplementary Material; adnominal possessives and numerals, as their treatment merits a full paper in itself.

We recast these adnominal word orders as comparative concepts, which are detailed in the Supplementary Material (Appendix D). There has recently been a lot of debate on the ontological nature of comparative concepts (see for instance *Linguistic Typology* 24.3); the six comparative concepts that we use in our case-study are of the 'hybrid type', in that they feature both semantic-functional and formal aspects (Haspelmath 2018: 86-87) and should be taken as cross-linguistically valid with respect to the eleven languages of our sample. Table 1 presents their implementation under our multi-layer approach, using two different components of UD and, where necessary, manually compiled list of lemmata.

COMPARATIVE CONCEPT	UD RELATION	UPOS	LIST OF LEMMATA
NOUN	-	NOUN, PROPN	-
RELATIVE CLAUSE	‘acl’ or ‘acl:relcl’	ADJ, NOUN, PROPN, VERB	-
MODIFYING ADJECTIVE	‘amod’	ADJ	-
ANALYTIC CASE MARKER	‘case’	ADP	adpositions
ARTICLE	‘det’	DET, PRON	articles
DEMONSTRATIVE	‘det’	DET, PRON	demonstratives

Table 1. A multi-layered implementation of the six comparative concept.

Both head and modifier can be restricted using different combinations of annotation layers (UD Relation, UPOS tags, lemmata), for a maximum of seven different combinations. As preliminary experimental work[15] has suggested that several combinations are in fact scarcely informative or redundant, we did not use all combinations of layers, instead treating corpus data as follows. First, we extract from CIEP all nominal heads and dependents using the relevant UD Relations; then, in order to mitigate parsing errors (noise), we further restrict data by filtering for relevant UPOS tags on both constituents. As an example, we present a snippet from our datafile in Figure 2; it lists nine instances of MODIFYING ADJECTIVE-NOUN dependencies from the first sentences of the English translation of García Márquez’s *Cien años de soledad*.



order	relfirst	relsecond	tokenfirst	tokensecond	lemmafirst	lemmasecond	uposfirst	upossecond	source
amod-X	amod	obj	firing	squad	fire	squad	VERB	NOUN	100YearsSolitude_English
amod-X	amod	obj	distant	afternoon	distant	afternoon	ADJ	NOUN	100YearsSolitude_English
amod-X	amod	nmod	clear	water	clear	water	ADJ	NOUN	100YearsSolitude_English
amod-X	amod	nmod	polished	stones	polished	stone	ADJ	NOUN	100YearsSolitude_English
amod-X	amod	obl	prehistoric	eggs	prehistoric	egg	ADJ	NOUN	100YearsSolitude_English
amod-X	amod	nsubj	many	things	many	thing	ADJ	NOUN	100YearsSolitude_English
amod-X	amod	nmod	ragged	gypsies	ragged	gypsy	ADJ	NOUN	100YearsSolitude_English
amod-X	amod	obl	great	uproar	great	uproar	ADJ	NOUN	100YearsSolitude_English
amod-X	amod	obj	new	inventions	new	invention	ADJ	NOUN	100YearsSolitude_English

Figure 2. A snippet from our datafile, showing the instances of modifying adjectives in the first lines of the English translation of García Márquez’s *Cien años de soledad*; cells highlighted in purple correspond to UD Relations and cells highlighted in yellow to UPOS.

These have been extracted by querying CIEP for the ‘amod’ UD Relation. If we subsequently apply the UPOS layer to both constituents, namely, exclude records that do not have the UPOS tag NOUN/PROPN for the head and that do not have the UPOS tag ADJ for the dependent, the first occurrence, *firing squad*, is excluded, as *firing* is not recognized as an adjective. However, note that our approach is not water-tight: note that *many things* is still recognized as an instance of MODIFYING ADJECTIVE, as the quantifier *many* is tagged as an adjective; the same for *polished stones*, as the participle *polished* is tagged as an adjective.

3.2 Results

3.2.1 Overview

In this first section, we will present an overview of all five word orders that were investigated: RELATIVE CLAUSE-NOUN, MODIFYING ADJECTIVE-NOUN, ANALYTIC CASE MARKER-NOUN, ARTICLE-NOUN and DEMONSTRATIVE-NOUN. Subsequent sections are devoted to considerations of each individual word order. We have plotted in Figure 3 the entropy values for all word orders. These are complemented by raw frequency data, discussed further on.

First of all, entropy values captured using only the UD Relations layer are higher, especially in the middle and lower part of the plot and for two comparative concepts, MODIFYING ADJECTIVE and ANALYTIC CASE MARKER. This indicates that the parser takes inappropriate parts of speech to be modifiers of a certain type (for example, a quantifier as a MODIFYING ADJECTIVE), or to be nominal heads (for example, a verb as nominal head). The introduction of this type of noise is common enough to affect entropy values for several word orders.

As for MODIFYING ADJECTIVE-NOUN order, the very high entropy amounts (>0.90) for the four Romance languages are not affected by using restrictions based on the additional UPOS layer, while a certain difference (0.1) can be observed for languages with medium values of entropy (0.60-0.70), Polish and Welsh, and low values of entropy (0.25-0.10), Dutch and German; the outlier here is however Danish, whose entropy drops from 0.64 to 0.25. As for ANALYTIC CASE MARKER-NOUN order, the restrictions based on the additional UPOS layer are particularly enlightening for the Germanic languages; with the exception of German, the entropy of these languages drops from 0.15-0.30 to a quasi-null value (<0.01).

Word order of the noun with modifiers RELATIVE CLAUSE, ARTICLE and DEMONSTRATIVE show low to very low amounts of entropy and very little difference when the restrictions based on the UPOS layer are introduced. Here we conflate ARTICLE and DEMONSTRATIVE into one category; these will be considered separately in Section 4.2.5. As we will show in Section 4.2.4 and Section 4.2.5, variation in the word order of ANALYTIC CASE MARKER, ARTICLE and DEMONSTRATIVE is better captured by combining UD Relations and UPOS with the manually-compiled lists of lemmata.

Entropy values can be compared with raw frequency data, which are given in Table 2 as frequency ratios and in Appendix E in the Supplementary Material as frequency values. Low variability (for example, analytic case markers in Dutch, which precede the noun in 95% of the attested dependency pairs) corresponds to low entropy (<0.10); high variability (for example Portuguese modifying adjectives, which precede the noun in 32% of the attested dependency pairs and follow it in 68%) corresponds to high entropy (>0.80). As suggested in Section 2.2, entropy can help capturing patterns of variation which would be otherwise hidden by a categorical classification relying on raw frequency data. If we were to apply Dryer's rule of thumb ("more than twice as common as the other order"), we should consider as dominant word orders with at least the .67 of frequency ratio; for instance, this would lead to classify postnominal modifying adjective as the dominant word order in Portuguese and Spanish, even though their entropy values are very high (>0.90).

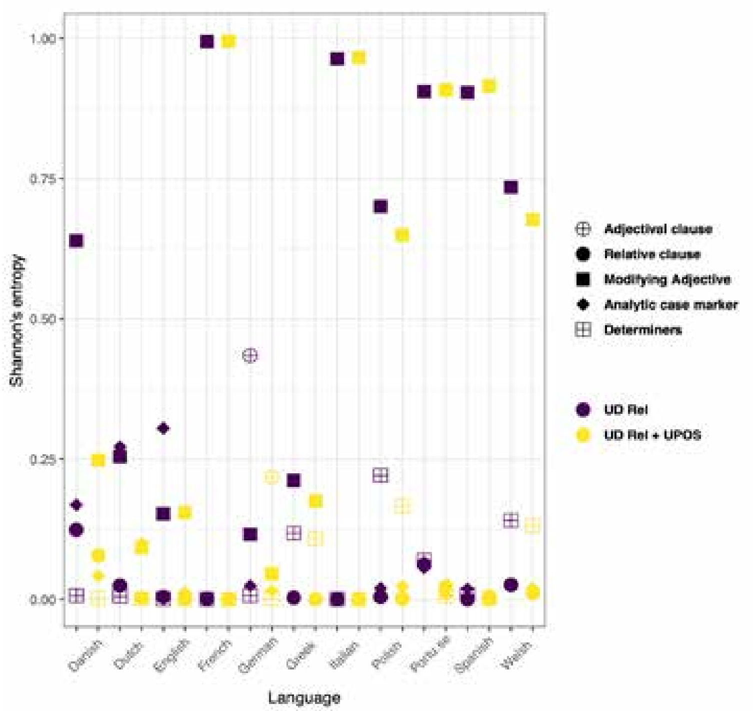


Figure 3. Entropy values for all comparative concepts using UD Relations for the modifier (purple) and UD Relations plus UPOS for both constituents (yellow); the adjectival clause is specific to German (see for more information Section 4.2.2 below).

In the remainder of this section, we will discuss all five word orders in detail; unless otherwise specified, we will use the combination of layers involving (a) the UD Relations and (b) restricting the data further based on the UPOS layer for both the head and the modifier as described above.

LANGUAGE	ACM		REL. CLAUSE		MOD. ADJECTIVE		**DETERMINERS**	
	h-m	**m-h**	**h-m**	**m-h**	**h-m**	**m-h**	**h-m**	**m-h**
Danish	.025	.975	.983	.017	.162	.838	.001	.999
Dutch	.047	.953	.998	.002	.043	.957	0	1
English	.054	.946	1	0	.022	.978	0	1
French	0	1	1	0	.545	.455	0	1
German	.002	998	n/a	n/a	.016	.984	.001	.999
Greek	0	1	1	0	.034	.966	.016	.984
Italian	0	1	1	0	.612	.388	0	1
Polish	.002	.998	1	0	.190	.810	.035	.965
Portuguese	.003	.997	.993	.007	.679	.321	.008	.992
Spanish	.002	.998	1	0	.681	.319	.001	.999
Welsh	.002	.998	.997	.003	.794	.206	.020	.980

Table 2. Frequency ratios for the head-modifier (h-m) and modifier-head (m-h) orders of three comparative concepts and the lexical category of determiners, using solely the UD Relations annotation layer.

3.2.2 Relative clause

The languages of the sample basically all have NOUN-RELATIVE CLAUSE order (Kurzová 1981). Exceptions are exceedingly rare, for example, Huddleston & Pullum (2002: 1066) cite one example of a 'preposed' relative clause in English, which is barely recognizable as a relative clause. Our data, which are plotted in Figure 4, confirm this assumption; however, there are two languages we need to consider further, German and Danish.

In German, attributive relative clauses are always introduced by a relative pronoun that inflects in number and gender with the head noun/pronoun, and they are always post-nominal (Breindl 2020e). In the German UD model (UD GSD v.2.5, see Appendix C in the Supplementary Material), there is no special tag for relative clauses,[16] as there is for the other ten languages of our sample. Rather, relative clauses fall into the general ‘acl’ relation, as described in Appendix D in the Supplementary Material; accordingly, we cannot produce reliable figures for relative clause in German, as the ‘acl’ relation captures all types of verbal modifiers.

Hawkins (1983: 13) states that “German has a minor prenominal relative clause strategy in addition to its postnominal relatives (*die den Mann liebende Frau* ‘the man loving woman’ / *die Frau, die den Mann liebt* ‘the woman who loves the man’)”. However, since the verb form of this supposed prenominal relative clause is a participle, and there is no relative pronoun, we would rather say that this is a more general verbal modifier, just like *das karottenliebende Pferd* ‘the carrot-loving horse’ or *das kürzlich entdeckte Fossil* ‘the recently discovered fossil’ (called a ‘modifying past participle’ by Holmberg & Rijkhoff 1998: 96-97). We include several of these examples from the corpus below. In all examples, modifiers as identified by the UD parser are indicated using the ‘acl’ and ‘amod’ Relation tags in superscript.

(1) N. Kazantzákis, Βίος και Πολιτεία του Αλέξη Ζορμπά (*Víos kai Politeía tou Aléxē Zorbá),*
German trans. by A. Steinmetz
Blaue Dämpfe stiegen aus dem Boden und verdichteten sich zu wechselnden$_{acl}$
blue vapor.PL rise.PST out.of the ground and condense.PST self to change.PTCP
Bildern: grinsenden$_{amod}$ *Mäulern, Füßen mit Krallen, die herannahten*$_{acl}$
image.PL grin.PTCP mouth.PL foot.PL with claw.PL REL get.closer.PTCP
und tief-schwarzen$_{amod}$ *Flügeln.*
and deep-black wing.PL
‘Blue vapors rose from the ground and condensed into changing images: grinning mouths, feet with claws, that were coming nearer, and jet black wings.’

(2) G. García Márquez, *Cien años de soledad,* German trans. by C. Meyer-Clason
Die Zigeuner-in entledigte sich ihrer über-einander-gezogenen$_{amod}$ *Mieder, ihrer*
the gypsy-F get.rid.PST.SG self her over-one.another-pull.PST.PTCP bodice.PL her
zahlreichen gestärkten$_{amod}$ *Spitzenröcke, ihres unnützen*$_{acl}$ *draht-verstärkten*$_{amod}$
numerous starch.PST.PTCP skirt.PL her unuse.PST.PTCP wire-reinforce.PST.PTCP
Korsetts, ihrer Glasperlen-last und stand plötzlich da, gewissermaßen
corset.PL her glass.bead.PL-load and stand.PST.SG suddenly there in.a.way
zu nichts verwandelt.
to nothing change.PST.PTCP
‘The gypsy got rid of several overlaying bodices, her numerous starched lace skirts, her useless wire-reinforced corset, her load of glass beads and suddenly stood there, in a sense transformed to nothing.’

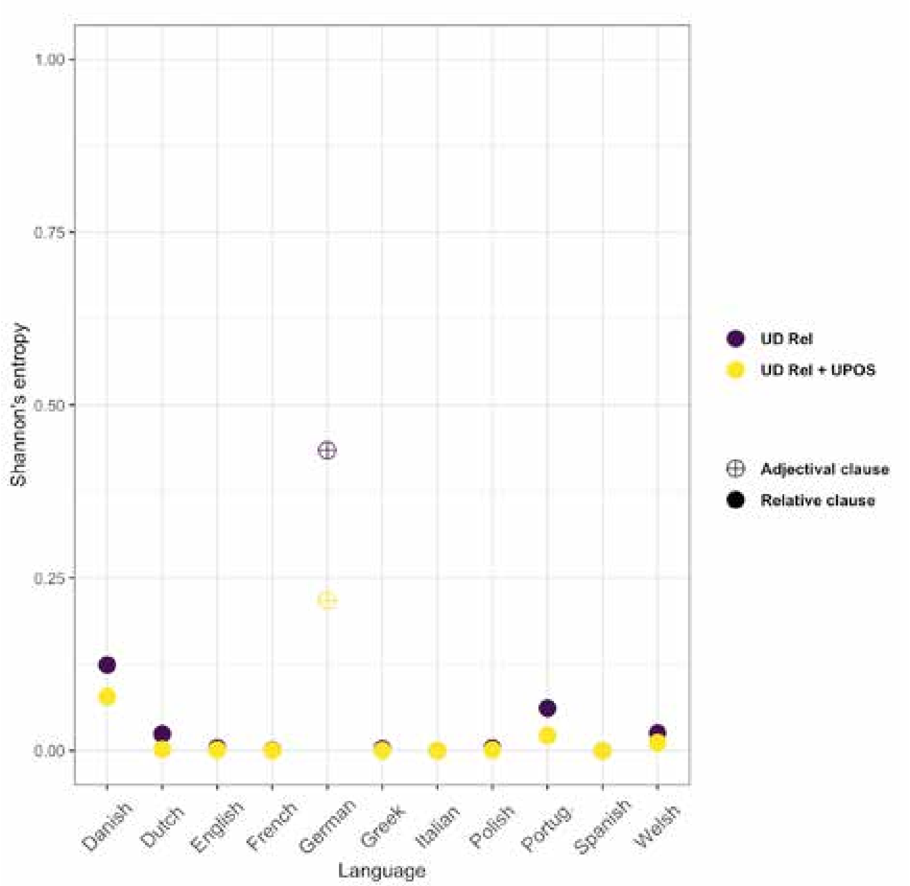


Figure 4. Entropy values for relative clause using UD Relations (purple) and UD Relations plus UPOS (yellow); the adjectival clause is specific to German.

We can observe from these examples that verbal modifiers in participial form are parsed both as ‘amod’ and as ‘acl’, without a clear pattern. Participal verbal modifiers (*ihres unnützen Korsetts*) are prenominal, true relative clauses are postnominal, collapsing the distinction between the two results in an entropy of 0.22 for adjectival clauses in German (prenominal ‘acl’: 539, postnominal ‘acl’: 15017). This raised entropy value is caused by a peculiarity of how relative clauses are treated in the German UD model.

Danish also has a non-zero entropy of relative clause and noun order, which is due to 200 occurrences of prenominal relative clause out a total of 20856. Lundskær-Nielsen & Holmes (2010: Section 4.8, 562-563) indicate that relative clauses are postnominal. However, the relative pronouns that introduce relative clauses (*der, som, hvad, hvem, hvis* and *hvilken* / *hvilke*t / *hvilk*e, Lundskær-Nielsen & Holmes 2010: 225) are homologous with other functional elements, similar to English *which* and *who*. *Hvad* is also an interrogative pronoun meaning ‘what’; its other forms are *hvem* ‘who(m)’ and *hvis* ‘whose, if’ (Lundskær-Nielsen & Holmes 2010: 215). *Hvilken* and its related forms *hvilket* and *hvilke* are also interrogative pronouns meaning ‘which, what’ (Lundskær-Nielsen & Holmes 2010: 215). *Som* is also a subordinator meaning ‘as’ (Lundskær-Nielsen & Holmes 2010: 525). And *der* is also an adverb meaning ‘there’, which is frequently used as an expletive subject. In addition, according to Lundskær-Nielsen & Holmes (2010: 230), *hvad* is used in non-restrictive relative clauses which may precede their head noun. However, there are no examples of this behavior presented in their grammar. The following are examples from CIEP where supposed but wrongly parsed prenominal relative clauses are identified using brackets:

(3) J. K. Rowling, *Harry Potter and the Philosopher’s Stone*, Danish trans. by H. Lützen

[Hvad stormhat og venus-vogn angår], er det blot to navne$_{head}$ for samme plante,
what stormhat and Venus-chariot concern.PRS is that/it just two name.PL for same plant
der beskytter mod vampyr-angreb; den går også under betegnelsen vinter-erantis.
that protect.PRS against vampire-attack it go.PRS also under term.DEF winter-Erantis
‘As for stormhat and Venus chariot, these are just two names for the same plant that protects against vampire attacks; it is also known as winter Erantis.’

(4) P. Coelho, *O Zahir*, Danish trans. by M. Hvass

[Politimanden vendte sig om mod butikkens indehaver]: – Hvis de
policeman.DEF turn.PST.SG self to against shop.DEF.GEN owner – if 2SG.FORM
får brug for det, er vi lige i nærheden$_{head}$.
get need for that is 1PL just in proximity
‘The policeman turned to the shop owner: – If you need it, we are close by.’

In both of these examples, there are multiple clauses which have a certain relation to each other; this has been interpreted by the parser such that one modifies the other, while this is not the case. In both cases, there is material between the supposed relative clause and its supposed head; this suggests that one way to further restrict noise like this is to not allow for anything to come between the relative clause and its head. In the first example, we are dealing with a free or ‘non-integrated’ relative clause (Huddleston & Pullum 2002: 1034), the type that we exclude here as it is not a nominal modifier. The second example is even more noisy, the second clause features *hvis* and we believe that might be why this analysis emerged, but why ‘*Hvis … nærheden*’ is parsed as the matrix clause is unclear.

We conclude that in Danish, prenominal relative clauses are introduced by the UD parser because of the homologous form of the relative pronouns. Why Danish would suffer more from this than other Standard Average European languages, which all have homologous relative pronouns to some extent, is unclear.

3.2.3 Modifying adjective

Out of all modifiers investigated in the current study, MODIFYING ADJECTIVE-NOUN order has the largest entropies, and also those which are most variable across languages (see Figure 5).

In Germanic languages and in Greek pre-nominal adjectives are the norm, but postnominal adjectives do occur, especially in contexts where modification is more extensive than one simple adjective or when the adjective is graded. Furthermore, postnominal adjectives may appear for semantic and pragmatic purposes, as in the case of the following Dutch example, in which the nature of the music was not previously established and is added as a note in passing.

In the following examples postnominal adjectival modifiers are marked with brackets:

(5) J. K. Rowling, *Harry Potter and the Half-Blood Prince*, Danish trans. by H. Lützen

(han)	*frigav*	*en*	*vandstråle*$_{head}$	*[så voldsom],*	*at*	*den*	*spulede*	*loftet*
3SG.M	release.PST	INDF	jet.of.water	so violent	that	DEF	splash.PST	ceiling.DEF
og fortsatte	*med*	*så stor*	*kraft ned*	*over professor Flitwick,*	*at*	*han*	*faldt*	
and continue.PST	with	so great	force down	over professor Flitwick	that	3SG.M	fall.PST	
fladt	*ned*	*på*	*maven.*					
flat	down	on	stomach					

‘(Lost in visions of this happy prospect, he flicked his wand a little too enthusiastically, so that instead of producing the fountain of pure water that was the object of today’s Charms lesson), he let out a hoselike jet that ricocheted off the ceiling and knocked Professor Flitwick flat on his face.’

(6) G. Musso, *La jeune fille et la nuit*, Dutch trans. by M. Meeuwes

De muziek$_{head}$	*– [gruwelijk en*	*meeslepend] –*	*begeleidde*	*je*	*tot op het centrale*
DEF music	– gruesome and	compel.PTCP –	accompany.PST	2SG	until on the central
plein	*van*	*het*	*lyceum.*		
square	of	the	lyceum		

‘The music – gruesome and compelling – accompanied one to the main square of the lyceum.’

(7) U. Eco, *Il nome della rosa*, German trans. by B. Kroeber

und zwischen der Oberlippe, die	*nicht*	*existierte,*	*und*	*der dicken, wulstigen*	*Unterlippe bleckten*		
and between the upper.lip REL	not	exist.AOR	and	the fat bulging	lower.lip bare.PST		
in unregelmäßigen	*Abständen*	*schwärzliche*	*Zähne*$_{head}$*,*	*[spitz*	*wie*	*die*	*eines*	*Hundes].*
in irregular	distance.PL	blackish	tooth.PL	sharp	like	REL	INDF	dog

‘… and between the upper lip, which did not exist, and the thick, bulging lower lip, blackish teeth, pointed like a dog’s, bared at irregular intervals.’

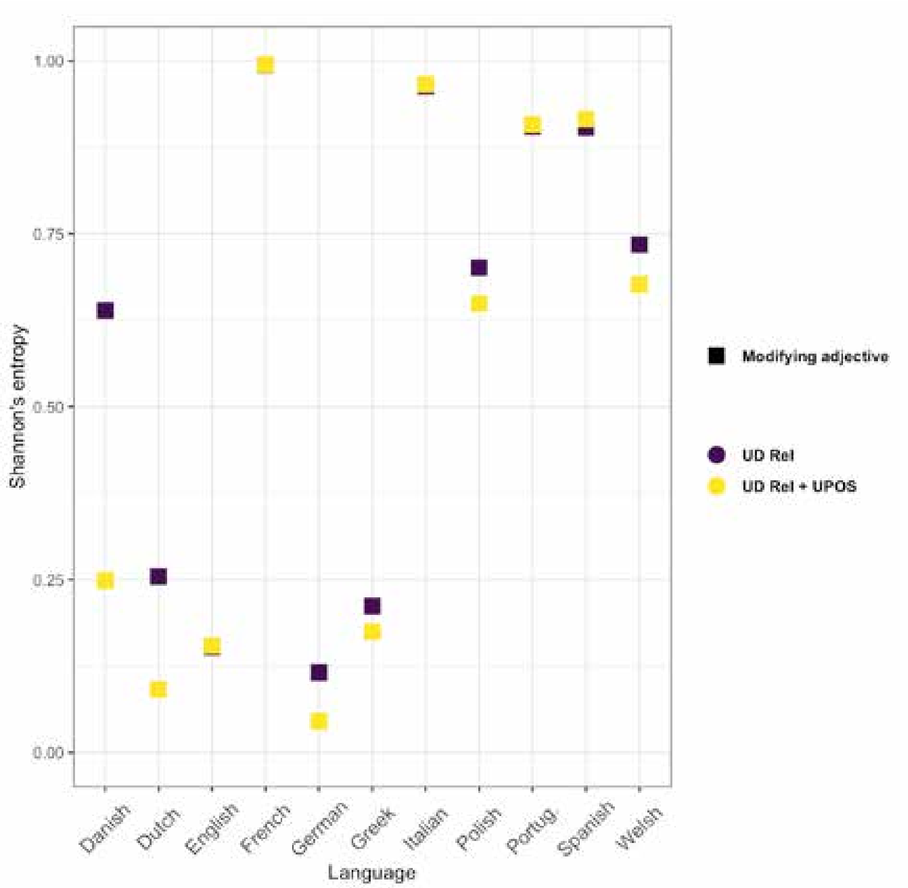


Figure 5. Entropy values for MODIFYING ADJECTIVE using UD Relations (purple) and UD Relations plus UPOS (yellow).

English examples:

- *They were hacking and stabbing at the ankles and shins of Death Eaters, their tiny faces$_{head}$ [alive with malice], …* (source: J. K. Rowling, Harry Potter and the Deathly Hallows);
- *It was shameful the way you left without compliments, as though she were an ancient hag$_{head}$ [a thousand years old].* (source: N. Kazantzakis, Βίος και Πολιτεία του Αλέξη Ζορμπά (*Víos kai Politeía tou Aléxē Zorbá*), English trans. by P. Bien);

- *…and a thousand more inventions*$_{head}$ *[so ingenious and unusual] that José Arcadio Buendía must have wanted to invent a memory machine so that he could remember them all.* (source: G. García Márquez, Cien años de soledad, English trans. by G. Rabassa)

Postnominal adjectives are possible in Greek by repeating the article in front of the adjective, as in τις φρέζιες τις κίτρινες / *tis frezies*$_{\text{head}}$ [*tis kitrines*] 'the yellow freesias (ACC)'; the construction has an emphatic reading (Lascaratou 1998: 164-165, Holton, Mackridge & Philippaki-Warburton 2012: 269). However, we were able to find only a few occurrences of this construction in our corpus, as in the following examples from *Het Achterhuis*:

(8) A. Frank, *Het Achterhuis*, Greek trans. by M. Grekou

a. και μετά πάλι πολύ εγωιστικά μόνο τις χαρές τις
kai metá páli polý egoistiká móno tis charés$_{head}$ *[tis*
and after again very selfishly only the.ACC.F.PL joy.ACC.(F).PL the.ACC.F.PL
δικές μου
dikés mou]
own.ACC.(F).PLmy
'[…] and again, in a very selfish way, my own joys […]'

b. το όνομα το χαϊδευτικό
to ónoma$_{head}$ *[to chaïdeftikó*]
the.NOM.N.SG name.NOM.(N).SG the.NOM.N.SG affectionate.NOM.N.SG
πατέρα μου
patéra mou
father my
'my father's affectionate name', i.e. 'my father's nickname'

The hundreds of occurrences (2276 out of 86926) of the noun-modifying adjective pattern are actually either instances of predicative adjectives or, worse, other parts of speech treated as adjectives; in fact, in our data the entropy of MODIFYING ADJECTIVE-NOUN order in Greek is close to zero.

In Welsh, the normal position of adjectives is postnominal; however, there are a number of prenominal adjectives, including *hen* 'old', *ambell* 'occasional', and *prif* 'main, chief'. Additionally, *pob* 'every, each' and *holl* 'all' are prenominal, these we would rather classify as quantifiers (King 2003: 69-70). Two examples from the corpus are included below, the first of a prenominal adjective (*hen* 'old'), and the second of a postnominal adjective (*bach* 'small'):

(9) J. K. Rowling, *Harry Potter and the Philosopher's Stone*, Welsh trans. by E. Huws Fel
Fel popeth arall oedd ganddo, bu'n berchen i aelod arall o'i
like everything other be.PST by.him be.PRF own.PST in member other of.his
deulu – ei daid yn yr achos yma. Ond doedd [hen] ddarnau$_{head}$
family – his granddad in DEF case here but be.NEG.3SG old piece.PL
gwyddbwyll ddim yn anfantais o gwbl.
chess not in disadvantage at all
'Like everything else he owned, it had once belonged to someone else in his family – in this case, his grandfather. However, old chessmen weren't a drawback at all.'

(10) L. Carroll, *Alice's Adventures in Wonderland*, Welsh trans. by S. Roberts
Llawdd iawn dweud "Yfwch Fi" ond roedd Alys$_{head}$ [bach] yn rhy gall i
easy very tell.INF drink me but be.PST.3SG Alice small in to can in
wneud hynny ar frys.
do that on hurry
'It was all very well to say 'Drink me,' but the (wise) little Alice was not going to do that in a hurry.'

The entropy is particularly high for Romance languages, with values exceeding 0.90. If we take a look at the raw frequency of MODIFYING ADJECTIVE in French, we can see a nearly equal division between the prenominal and the postnominal positions: 37340 prenominal *vs* 44226 postnominal. The ratio in the other Romance languages is shifted more towards the postnominal position, but again the position of the MODIFYING ADJECTIVE is variable: Italian – 30033 prenominal *vs* 46633 postnominal, Portuguese – 21530 prenominal *vs* 45036 postnominal, Spanish – 22976 prenominal *vs* 46489 postnominal.

This variability is somewhat reflected in the confusing and often contradictory treatment of the position of MODIFYING ADJECTIVE in grammars of Romance languages. For instance, a French reference grammar offers a list of "adjectives which normally precede or follow the noun", "adjectives which change their meaning according to their position" and "adjectives whose position does not affect meaning" in the form of types of adjectives such as 'short, very common adjective', 'color', 'nationality', and so on (Batchelor & Chebli-Saadi 2011: 663-666); a very similar list is offered by a Brazilian Portuguese reference grammar, both in the form of types of adjectives, 'ordinal numbers' and 'indefinite/quantitative adjectives' and in individual adjectives (Whitlam 2011: 38-41). However, both the French and the Brazilian Portuguese grammars also write that '[on the basis of register and usage] there is an increasing tendency to place adjectives which normally follow the noun before it" (Batchelor & Chebli-Saadi 2011: 664) and "the default position for attributive adjectives is after the noun in Portuguese [...] however, the rule can be broken when an adjective is used not to differentiate or specify, but rather to mention an inherent quality of the noun. This kind of stylistic device is mainly confined to the written language, especially journalistic style" (Whitlam 2011: 38).

For the other two Romance languages, Italian and Spanish, consulted grammars seem to offer a more reassuring scenario; in both languages, the prenominal position is the normal position, while the postnominal position would have a delimiting function, namely “serve to identify, pick out, highlight, place in the foreground, focus attention on, a subset of the entities referred to by the noun” (Italian: Maiden & Robustelli 2013: 48) and “narrow the scope of the noun that precedes them” (Spanish: Butt, Benjamin & Rodríguez 2019: 65). Authors like Cinque (2010) claim that the prenominal position of Italian adjectives has a specific meaning, such as restrictive, which excludes a non-restrictive sense (this applies to other meanings as well, only one reading is possible prenominally), while postnominal Italian adjectives would be ambiguous, in the sense that they could be restrictive or non-restrictive (as well as take both values on other oppositions). This characterization is similar to the above-mentioned delimiting function attributed to prenominal adjectives. Similar claims can be extended to the position of the Spanish adjective, as the Spanish grammar attributes to postnominal adjectives an ambiguous reading between non-restrictive and restrictive values (Butt, Benjamin & Rodríguez 2019: 65); however, as their French and Portuguese counterparts, the Spanish grammarians note that “Unfortunately the distinction between restrictive and non-restrictive adjectives is not always clear, so the decision about where to put the adjective sometimes relies on a feel for the language rare among non-natives” (Butt, Benjamin & Rodríguez 2019: 65).

Such “feel for the language” is a suggestive term for the deep mastery of the language’s pragmatics and lexicon; it seems, then, that in the position of Romance MODIFYING ADJECTIVES, pragmatic and even stylistic factors intervene on an already complicated and lexically governed situation. To exemplify this matter, take the following parallel sentences from the translations of Süskind’s *Das Parfum*; as in the examples above, adjectival modifiers are marked with brackets:

(11)

a. P. Süskind, *Das Parfum*, French trans. by B. Lortholary
Il *tira* *un [petit]* *mouchoir*$_{head}$ *de* *dentelle,* *[frais* *et*
he pull.PST a little.M.SG handkerchief.(M).SG of lace fresh.M.SG and
blanc] *comme neige,* *de la poche* *de son habit,* *de la poche*$_{head}$
white.M.SG like snow from the pocket of his suit from the pocket.(F).SG
[gauche], *le déploya* *et* *y* *fit* *tomber* *quelques gouttes* *puisées* *dans la*
left.F.SG it unfold.PST and LOC let drop few drops drawn from the
bouteille *à mélanger* *avec la [longue] pipette*$_{head}$.
bottle to mix with the long.F.SG pipette.(F).SG
‘He pulled a small lace handkerchief, fresh and white as snow, from the pocket of his suit, from the left pocket, unfolded it and dropped a few drops from the bottle to mix with the long pipette.’

b. P. Süskind, *Das Parfum*, Italian trans. by G. Agabio
Prese dalla tasca$_{head}$ *della giacca, dalla [sinistra], un*
take.PST from_the pocket.(F).SG of_the jacket from_the left.F.SG a
fazzoletto$_{head}$ *[pulito] di pizzo, [bianco] come la neve, lo*
handkerchief.(M).SG clean.M.SG of lace white.M.SG like the snow it
spiegò e vi spruzzò sopra un paio di gocce prese con la
unfold.PST and LOC spray.PST on a couple of drops drawn with the
pipetta$_{head}$ *[lunga] dalla bottiglia della miscela.*
pipette.(F).SG long.F.SG from_the bottle of.the mixture
'He took from his jacket pocket, from the left one, a clean lace handkerchief, white as snow, unfolded it and sprayed on it a couple of drops taken with the long pipette from the mixture bottle.'

c. P. Süskind, *Das Parfum*, Portuguese trans. by F. R. Kothe
Puxou um lenço$_{head}$ *de renda, bem [fresco e limpo], do*
pull.PST a handkerchief.(M).SG of lace very fresh.M.SG and clean.(M).SG of
bolso$_{head}$ *[esquerdo] do jaquetão, desdobrou-o e borrifou sobre ele*
pocket.(M).SG left.M.SG of jacket unfolded-it and spray.PST on it
algumas gotas que com a pipeta$_{head}$ *[longa] extraíra da garrafa.*
some drops that with a pipette.(F).SG long.F.SG extract.PST from bottle
'He pulled a very fresh and clean lace handkerchief from the left pocket of his jacket, unfolded it, and sprayed on it a few drops that he had extracted from the bottle with a long pipette.'

d. P. Süskind, *Das Parfum*, Spanish trans. by P. Giralt Gorina
Extrajo del bolsillo$_{head}$ *[izquierdo] de la levita un [pequeño]*
take.PST from_the pocket.(M).SG left.M.SG from the coat a small.M.SG
pañuelo$_{head}$ *de encaje*$_{head}$ *[blanco] como la nieve, lo desdobló y*
handkerchief.M.SG of lace.(M).SG white.M.SG like the snow it unfold.PST and
lo humedeció con un par de gotas que sacó del matraz mediante
it moisten.PST with a couple of drops that remove.PST of_the bottle with
la [larga] pipeta$_{head}$.
the large.F.SG pipette.(F).SG
'He took a small snow-white lace handkerchief out of his left coat pocket, unfolded it and moistened it with a couple of drops, which he removed from the flask with the long pipette.'

The four Romance languages concord on the positions of the MODIFYING ADJECTIVE whose nominal head is 'pocket', which happens to be postnominal in all languages; hence, the MODIFYING ADJECTIVE 'left' would be ambiguous between a restrictive and non-restrictive reading. However, the different translations are more prone to a restrictive reading; Italian uses a headless, anaphoric adjective, while French repeats the nominal head. There is also concordance between French and Spanish on the position of the MODIFYING ADJECTIVE 'small', which is placed in the prenominal, hence restrictive, position; however, rather than pointing to a specific handkerchief, the two lexemes *petit/pequeño* 'small' denote it for its intrinsic property of being small, translating the original, non-modified lexeme *Spitzentüchlein* ‘small lace handkerchief’. Furthermore, the original nominal phrase *der langen Pipette* ‘the long pipette’ is translated into the four Romance languages by using both strategies, with French and Spanish using the prenominal position and Italian and Portuguese the postnominal position, hence testifying once again the inter- and intra-linguistic variability of the position of Romance MODIFYING ADJECTIVES.

Finally, the only representative in our sample of a Slavic language, Polish, shows a moderately high entropy, 0.65 (prenominal adjectives: 67669, postnominal adjectives: 13501); according to Siewierska & Uhlířová (1998: 134), the position of adjectives in Polish is determined by their lexical type, with qualitative or evaluative adjectives favoring the prenominal position and relational or denominal adjectives the postnominal position. Our data confirms Siewierska & Uhlířová’s claim: among the most frequent adjectives in prenominal position we find qualitative lexemes like *piękny* ‘beautiful’ (267 occurrences) and *długi* ‘long’ (854 occurrences), while relational adjectives[17] such as *domowy* ‘domestic’ (209 occurrences) and *wejściowy* ‘relating to entrance’ (111 occurrences) figure in the first frequency ranks of postnominal adjectives. However, as already noted by Siewierska & Uhlířová (1998: 134), the position of the modifying adjective can be reversed for ‘reasons of focus or emphasis’; for instance, we find 20 occurrences of qualitative *piękny* and 28 occurrences of qualitative *długi* in the postnominal position, and 110 occurrences of relational *domowy* and 5 occurrences of relational *wejściowy* in the prenominal position. The following passage from the translation of Eco’s *Il nome della rosa* contains an example of *długi* in the postnominal position, where particular emphasis is placed on the length of the fingers’ character:

(12) U. Eco, *Il Nome della rosa*, Polish trans. by A. Szymanowski

Dłonie	*miał*	*białe,*	*palce*	*długie*	*i*	*szczupłe.*
hands	have.PST.3SG.M	white	fingers	long	and	thin

‘He had white hands, long and thin fingers.’

3.2.4 Analytic case marker

Figure 6 depicts the entropy of the order of ANALYTIC CASE MARKER and NOUN, introducing an additional annotation layer, Lists of Lemmata (LoL). As discussed above, LoL are manually created list of words curated from grammars (see Appendix F in the Supplementary Material), usually belonging to a closed lexical category. We observe that filtering query results using lists of lemmata reduces entropy by getting rid of noise introduced by the parser, dropping the entropy values to very low levels (< 0.05) in all languages but one.

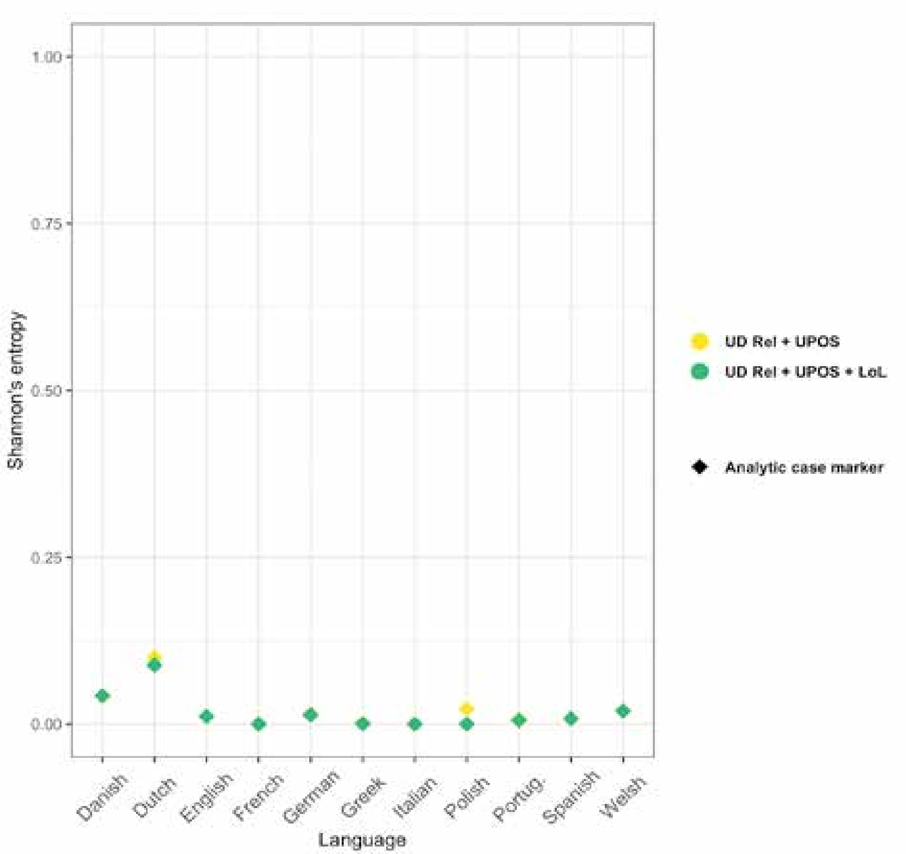


Figure 6. Entropy values for analytic case marker, using the combination between UD Relations and UPOS (yellow) and the combination between UD Relations, UPOS and lists of lemmata (green).

This suggests there is little to no variation and the languages of the sample are mainly prepositional; however, several languages have a minority lexical category of postpositions, ambipositions, or circumpositions:

(i) Danish: circumpositions: *ad … til, for … siden, fra … af, på … nær, for … skyld, i … sted, på … vergne*; 'postposed prepositions': *foruden, igennem, over* (Lundskær-Nielsen & Holmes 2010: 424-425, see also Hagège 2010: 126)

(ii) Dutch: *aangaande, af, binnen, door, in, langs, niettegenstaande, om, op, over, rond, uit, uitgezonderd, voorbij* (Hagège 2010: 119-121, Broekhuis 2013: 33-34; see Broekhuis, 2013: 48-66 for a discussion and classification of Dutch circumpositions)

(iii) French: *comprises, durant, exceptées* (Hagège 2010: 119)

(iv) English: *ago, apart, aside, notwithstanding, on* (Huddleston & Pullum 2002: 631-632)

(v) German: *ausgenommen, bar, betreffend, entgegen, entlang, entsprechend, gegenüber, gemäß, inbegriffen, nach, nahe, um … willen, unbeschadet, ungeachtet, wegen, zufolge, zuliebe, zunächst* (Breindl 2020d, Hagège 2010: 121)

(vi) Greek: ἕνεκεν, χάριν (Holton, Mackridge & Philippaki-Warburton 2012: 498)

(vii) Italian,[18] Polish (Siewierska & Uhlířová 1998:110), Portuguese, Spanish and Welsh do not seem to have postpositions.

Aside from these minority patterns, which occur with particular adpositions or in specific contexts, Dutch has the only entropy value that convincingly deviates from zero; in Dutch, we find 1704 occurrences of postnominal analytic case marker out a total number of 152,629, resulting in an entropy value of 0.09. This is due to a phenomenon called *R-pronominalization* (Broekhuis 2013: Chapter 5) or *EDH-postpositions* (Berendsen 2021). In prepositional phrases where the complement of the preposition is a pronoun *het* 'it', or a demonstrative *dit* 'this', *deze* 'these', *dat* 'that', or *die* 'those', the pronouns are replaced by *er, hier* (proximate demonstratives) and *daar* (for the distal demonstratives). The preposition is moved to a position after *er, hier*, or *daar* (EHD-words), hence creating postpositions, or rather, postpositional use of prepositions. Examples from the corpus with *er* and *daar* follow in examples (13) and (15) with the corresponding clause without *er* or *daar* in examples (14) and (16). In these examples, the analytic case markers are given in deitalicized script, and the head nouns are put in between brackets.

(13) J. K. Rowling, *Harry Potter and the Chamber of Secrets*, Dutch trans. by W. Buddingh'
Hedwig en Egidius *keken* *geïnteresseerd toe terwijl Harry het*
Hedwig and Egidius look.PST.3PL interest.PTCP to while Harry DEF
tegenspartelende boek *tegen zijn borst klemde,* *haastig naar zijn ladekas*t
struggle.ADJ book against his chest pin.PST.3SG hurriedly to his drawers
liep, *[er]* *een* *riem* uit *haalde* *en die strak* *om het boek*
walk.PST.3PL ER INDF belt out take.PST.SG and that tight around DEF book
gespte.
buckle.PST.SG
'Hedwig and Egidius looked on interestedly while Harry pressed the struggling book against his chest, hurried to his drawers, took a belt out and buckled it tight around the book.'

(14) Parallel of (13) with non-pronominal complement of *uit*
Harry haalde *een* *riem* uit *[zijn ladekast].*
Harry take.PST.3SG INDF belt out his drawers.
'Harry took a belt out of his drawers.'

(15) J. K. Rowling, *Harry Potter and the Chamber of Secrets*, Dutch translation by W. Buddingh'
[Daar] *heb* *je* *gewoon niet genoeg* *tijd* voor.
there have.PST.2SG you simply not enough time for
'You simply don't have enough time for that.'

(16) Parallel of (15) with non-pronominal complement of *voor*:
Je *hebt* *gewoon niet* *genoeg* *tijd* voor *[…].*
you have.PST.2SG simply not enough time for […]
'You simply don't have enough time for …'

As is evident from the examples in (13-16), while using EHD-words is obligatory if the nominal complement of a preposition is pronominalized, using EHD-words and EHD-postpositions is also a convenient way to manipulate information structure through word order, for instance fronting the pronoun in example (15), or referring to an entity already established in example (13). While there is considerable noise in correctly identifying EHD-postpositions in CIEP through the UD parser, it is a frequent phenomenon and the reason for the non-zero entropy of the order of analytic case marker and noun for Dutch.

3.2.5 Article and demonstrative

As detailed in Appendix D in the Supplementary Material, the UPOS DET and the UD Relation 'det' cover a wide array of word categories such as articles, demonstratives, quantifiers, interrogative and personal/possessive pronouns, which are collectively treated as 'determiners'; we have chosen to restrict the category of determiners to the smaller and more comparable categories of DEMONSTRATIVE and ARTICLE. The orders of DEMONSTRATIVE-NOUN and ARTICLE-NOUN are captured by the combination between UD Relations, restrictions based on the UPOS and LoL, while the determiner-NOUN order is investigated on the combination using the UD Relation 'det' and restrictions on the head noun based on UPOS tag NOUN. In Figure 7 we present entropy values for the word order of DEMONSTRATIVE and NOUN, ARTICLE and NOUN, and determiner and NOUN.

There is no word order variation in the position of the ARTICLE; a more interesting scenario is offered by the DEMONSTRATIVE, for which we find a high entropy in Greek (0.73) and a moderate entropy in Welsh (0.37); furthermore, a low, but still raised, entropy is also attested in Polish (0.15).

In Greek, demonstratives co-occur with definite article and are normally placed prenominally; however, they can be placed postnominally 'for reasons of emphasis' (Lascaratou 1998:164). We find 1448 postnominal demonstratives out a total of 7012. As with the Romance and Slavic MODIFYING ADJECTIVE, this variation in the position of Greek demonstratives has a semantic and/or pragmatic function; for instance, in example (17), αυτή *aftí* 'this' is placed after the nominal head είδηση *eídisi* 'news' in order to stress how the news would have caused discomfort to the subject's parents; similarly, in example (18), εκείνη *ekeíni* 'that' highlights the night (νύχτα *nýchta*) in which the capture took place.

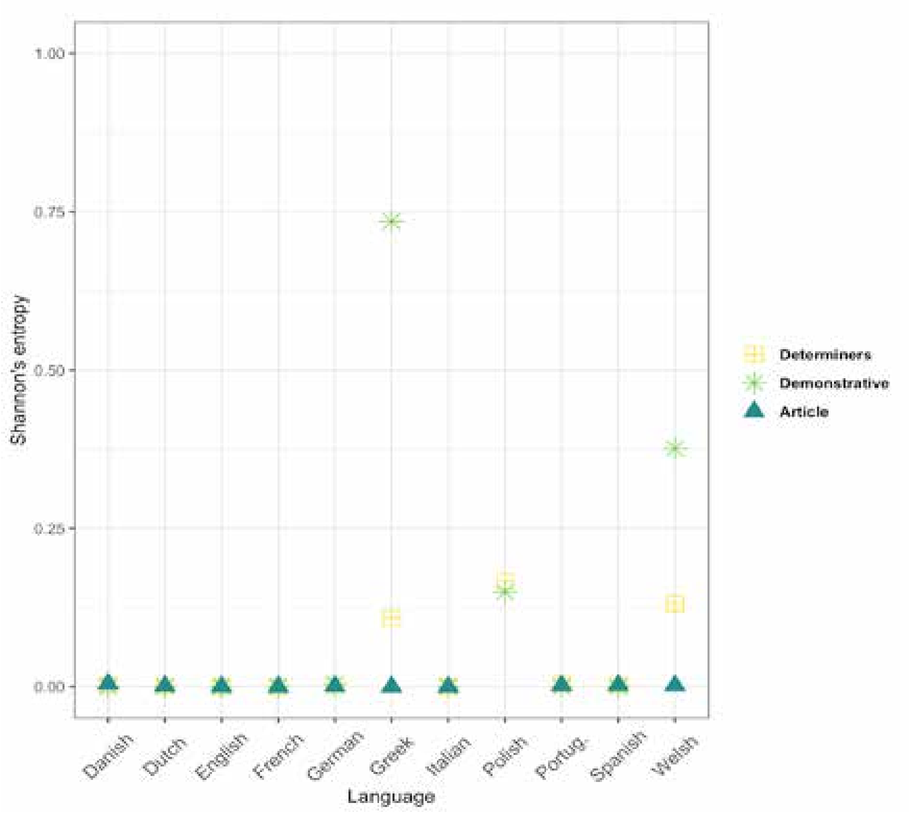


Figure 7. Entropy values for demonstrative (light green) and article (light blue), using the combination between UD Relations, UPOS and LoL. The lexical category of determiners uses the combination between UD Relations and UPOS (yellow).

(16) G. García Márquez, *Cien años de soledad*, Greek trans. by M. Palaiologou
εκείνος ένιωσε πως η είδηση αυτή θα ήταν η χαριστική
ekeínos éniose pos i eídisi aftí tha ítan i charistikí
he feel.AOR.3SG how the.F.SG news.(F).SG this.F.SG COND be.PST the gratuitous
βολή για τους γονείς του
volí gia tous goneís tou
blow for the parents his
'He felt that this news would be the final blow for his parents.'

(17) J. K. Rowling, *Harry Potter and the Chamber of Secrets*, Greek trans. by K. Oikonomou
θα σε πάρω μέσα στις αναμνήσεις μου τη νύχτα εκείνη
tha se páro mesa stis anamníseis mou ti nýchta ekeíni
FUT you.ACC take into in.the memories my the.F.SG night.(F).SG that.F.SG
που τον έπιασα
pou ton épiasa
where him catch.AOR.1SG
'I'll take you into my memories of the night I caught him.'

The co-occurrence of articles and demonstratives is also attested in Welsh, when the demonstrative is built through a construction with a prenominal article and a postnominal demonstrative (King 2003: 29, 85), causing determiner-noun entropy to be 0.13 (see Figure 7). While Welsh articles show no word order variation, demonstratives display a moderate entropy, 0.37. This is caused by the highly limited number of occurrences of adnominal demonstratives in the corpus. Welsh demonstratives are *hwn*, *hwnnw*, *hon*, *honno*, *hyn*, and *hynny* (King 2003: 85). Adverbials used as demonstratives are *na* and *ma* (which mean 'here' and 'there', King 2003: 85). The later category is excluded here, as these are analyzed as ADP in UD UPOS, and their associated UD Relation is 'case'. The true demonstratives appear frequently in the corpus, but not often as nominal modifiers. There are 21 instances of *hwn, wwnnw*, etc. as prenominal modifiers, and 268 instances as postnominal modifiers. Obviously the Welsh subcorpus is smaller than all other subcorpora (see Section 3 and Appendix C in the Supplementary Material), but this cannot be the full explanation. We blame the infrequent use of nominal modifier demonstratives to the usage of *'na* and *'ma* as demonstratives, which we exclude here, and additional factors we do not currently understand.

The demonstrative in Polish is generally prenominal, but it can also occur in the postnominal position; according to Siewierska & Uhlířová (1998: 132-133), postnominal demonstratives are limited to the proximate demonstrative and their placement has a discourse (anaphoric) function, as in example (18), in which the postnominal demonstrative in *spraw tych* '(of) these matters' anaphorically refers to the list of matters introduced in the previous clause.

(18) P. Süskind, *Das Parfum*, Polish trans. by M. Łukasiewicz
Wprawdzie nie posunął się do tego, by – jak to
although not go.PRF REFL to this.PROX.GEN.M.SG to – as this.PROX.NOM.N.SG
się zdarzało niektórym – kwestionować cuda, proroctwa albo prawdę
REFL happen.IPFV.PST some.DAT – question.IPFV.INF miracles prophecies or truth
Pisma Świętego, jakkolwiek ściśle rzecz biorąc, spraw
scripture holy however strictly thing taking matter.GEN.PL.(F)
tych nie dawało się wyjaśnić wyłącznie rozumowo […]
this.PROX.GEN.PL.F not give.IPFV REFL explain.PFV.INF only reason.ADV
'Although he did not go so far as to question miracles, prophecies, or the truth of Holy Scriptures, as some have done, however strictly you take the thing, these matters could not be explained by reason alone […]'

Out of 11257 occurrences of demonstrative, only 242 are postnominal and the great majority (218) is represented by *ten* 'this (citation form)'; the low entropy value (0.15) of Polish demonstrative-noun order is explained by the text genre peculiar to CIEP, that is, prose; according to Siewierska & Uhlířová (1998: 133), the postnominal strategy of Polish demonstrative "is characteristic of the written language, particularly of expository and journalistic texts".

4. General discussion

We believe that it is surprising how much word order variation we find in our case-study, considering how prevalent categorical measures in typology are, and considering that some (but not all) cases of variation described are well-known in typology. For summary purposes, Table 3 presents the entropy of the five word orders investigated in the previous section. Table 3 should be compared with Table 4, which shows the classification of WALS (Dryer & Haspelmath 2013) and various chapters from Siewierska (1998).[19]

Through the comparison of Tables 3 and 4 we can see that the binary classification of NOUN-RELATIVE CLAUSE and ARTICLE-NOUN order in WALS is an appropriate classification for these eleven languages. However, this is not the case for the order of ANALYTIC CASE MARKER, MODIFYING ADJECTIVE, and DEMONSTRATIVE with respect to the NOUN. The latter three show significant variation in modifier-head word order, for the ANALYTIC CASE MARKER this concerns only Dutch, but for MODIFYING ADJECTIVE and DEMONSTRATIVE variation is spread across more languages, and for different reasons, as detailed in Section 4.2.

LANGUAGE	ACM	REL. CLAUSE	MOD. ADJECTIVE	ARTICLE	DEMONSTRATIVE
Danish	0.04	0.08	0.25	0	0
Dutch	0.10	0	0.09	0	0
English	0.01	0	0.15	0	0
French	0	0	0.99	0	0
German	0.01	n/a	0.04	0	0
Greek	0	0	0.17	0	0.73
Italian	0	0	0.96	0	0
Polish	0	0	0.65	-	0.15
Portuguese	0	0.02	0.90	0	0
Spanish	0	0	0.91	0	0
Welsh	0.02	0.01	0.68	0	0.37

Table 3. Entropy values for the order of five modifier-noun word orders considered in the present study; we use the combination between UD Relations and UPOS for MODIFYING ADJECTIVE and RELATIVE CLAUSE, and the combination between UD Relations, UPOS and LoL for ANALYTIC CASE MARKER (ACM), ARTICLE and DEMONSTRATIVE.

We can compare our entropy scores with those by Levshina (2019), especially her first set of case studies, in which she conducts analyses of cross-linguistic and intra-linguistic variability (Levshina 2019: 542-551; see the table in Appendix A, Supplementary Material). In her analysis, determiner-noun word order (det_Noun) is associated with both low cross-linguistic and intra-linguistic variability, whereas adposition-noun order (adp_Noun), adjective-noun order (amod_Noun), and adjectival clause-noun order (acl_Noun) are associated with high cross-linguistic variability, but low intra-linguistic variability. This matches our results only partially. We observe more word order variability for DEMONSTRATIVE-NOUN order, but only for a few languages (Greek, Polish, and Welsh). Our scores and Levshina's are however not directly comparable here, as DEMONSTRATIVE-NOUN order only covers a part of Levshina's det_Noun and probably represents the word category with the highest variability among determiners. A more readily comparable word order is instead the one involving MODIFYING ADJECTIVE; when considering only the UD Relations annotation layer, this corresponds to Levshina's amod_Noun. Here we observe much more cross-linguistic and intra-linguistic variability than attested in Levshina (2019). We have found a particularly high variability for Romance languages and high variability for Polish and Welsh, which contrasts with the lower entropy attested in the Germanic languages and Greek.

LANGUAGE	ADPOSITION & NOUN	RELATIVE CLAUSE & NOUN	ADJECTIVE & NOUN	ARTICLE & NOUN	DEMONSTRATIVE & NOUN
Danish	adposition-noun	noun-relative clause	adjective-noun	article-noun	demonstrative-noun
Dutch	adposition-noun	noun-relative clause	adjective-noun	article-noun	demonstrative-noun
English	adposition-noun	noun-relative clause	adjective-noun	article-noun	demonstrative-noun
French	adposition-noun	noun-relative clause	noun-adjective	article-noun	demonstrative-noun
German	adposition-noun	noun-relative clause	adjective-noun	article-noun	demonstrative-noun
Greek	adposition-noun	noun-relative clause	adjective-noun	article-noun	demonstrative-noun
Italian	adposition-noun	noun-relative clause	noun-adjective	article-noun	demonstrative-noun
Polish	adposition-noun	noun-relative clause	adjective-noun	-	demonstrative-noun
Portuguese	adposition-noun	noun-relative clause	noun-adjective	article-noun	demonstrative-noun
Spanish	adposition-noun	noun-relative clause	noun-adjective	article-noun	demonstrative-noun
Welsh	adposition-noun	noun-relative clause	noun-adjective	article-noun	noun-demonstrative

Table 4. WALS's classification for the word orders considered in the present study; data for Article & Noun comes from individual chapters in Siewierska (1998).

Differences between Levshina's (2019) entropy scores and ours may be explained by the different text genres used as data. For her first set of case studies, Levshina uses the UD treebanks which, with respect to the subset of CIEP used in the case-study, mostly feature texts of the news, non-fiction and encyclopedic (Wikipedia) genres.[20] As mentioned in Section 4.2, the UD parsers introduce noise. However, a case-study in Levshina *et al. to appear* reports that the difference between the entropies of the same four nominal modifiers in CIEP and UD treebanks is not statistically significant. This implies that CIEP, as a corpus of fiction, seems to allow higher variability in the order of DEMONSTRATIVE-NOUN and MODIFYING ADJECTIVE-NOUN.

So far, studies on word order variability have used almost exclusively UD treebanks (Naranjo & Becker 2018, Alzetta *et al.* 2018, Gerdes, Kahane & Chen 2019) or similar projects (HamleDT: Futrell, Mahowald & Gibson 2015). We are still standing at the cradle of token-based typology, and as of yet we do not know much about the effect of source material on the corpus-based analysis of word order or of other typological features (there is very little to cite here: see Levshina 2015 for a highly reasoned choice for a parallel corpus of film subtitles to study forms of address). It is a topic of further investigation which type of source material would fit best to investigate different typological questions. However, we would argue in general for a closer consideration of (sub)corpus when choosing cross-linguistic materials. UD treebanks dramatically differ across languages with respect to composition (and size). Analyzing these materials as given, as is often done, considers all this material as a single undifferentiated unit representing 'Language X', while in fact the feature at hand may show different behavior across registers. This may be the case for MODIFYING ADJECTIVE-NOUN order in the current study, and has been described at length for English and other languages by Biber (1993), Biber (1995), and Biber (2012). As described in Section 3, CIEP contains only fiction, so only one register, but it does represent variation within that domain, i.e. novels from different times and places, different styles, and different 'register-within-registers' such as dialog (some of it quite naturalistic), diary entries, and letters. This subgenre variation can be explored in future work.

In evaluating our case-study, we see the parallel nature of CIEP as its major benefit: differences or similarities that we find between languages can be directly linked to linguistic phenomena. When using non-parallel alternatives such as the UD treebanks, the different nature of the material of each individual treebank (both in terms of register and content) may influence the results. We believe this is especially important when investigating complex phenomena such as word order variability, which we show here to be not only dependent on grammatical rules ('In Dutch, the adposition stands before the noun') but also on information structure and pragmatics.[21]

However, the choice of a parallel corpus of fiction as a data source does not come without issues; while we see the benefits of having translational equivalents, others have argued that linguistic phenomena in translation may be influenced by the original text (Santos 1995, Gellerstam 1996, Altenberg & Granger 2002, Cappelle 2012) or are otherwise biased as translated rather than original language use (Wälchli 2007). While this topic features largely in corpus studies and contrastive studies, it is unclear what the impact of the translation bias is (a) across (typologically different) languages and (b) for different typological features of interest. In the current case of word order variability, it seems likely that languages with less strict grammatical rules would suffer more (i.e. adopt the original order, even though it might not be native-like). However, since word order in those languages is typically governed by information structure and pragmatics, which would take precedence over keeping the original order, this prediction might be untrue. In short, we do not know yet enough about how the translation process might impact word order research using token-based typology, let alone the investigation of different typological features.

Another issue with using CIEP is that we use UD parsers rather than treebanks, which are at least partly annotated by humans. This introduces a certain amount of noise in our data, especially for low-resource languages such as Welsh and, to a minor extent, Greek. Furthermore, as is evident from the table in Appendix C (Supplementary Material), most of the models have been trained on other text genres than fiction, with a strong bias on web varieties, i.e. blog and wiki. We essentially face here technical issues and scarcity of resources, which will hopefully be resolved with the advancement of the UD project and, more in general, of cross-linguistic NLP tools. However, we have at least partly solved this issue by working with the UD parsers rather than using them as given. We have approached them with valid cross-linguistic comparative concepts in mind and matched these to the relevant UD Relations and Part of Speech tagging. This is a procedure we recommend for those that venture to use UD treebanks or parsers, to consider ways in which to tune the UD annotation and parsing in such a way to better match the phenomenon under analysis.

We would like to end this paper with a viewpoint on the word order variability that we have demonstrated. Capturing word order not only in categorical classifications but also in continuous classifications, including the entropy measure used in the current study, is of great importance for accurate measurement on several levels, including diachrony, universals, and explanations of typological distributions.

First, let us consider diachronic typology again. If we wanted to explain the supposed NOUN-MODIFYING ADJECTIVE word order of the Romance languages in terms of a rigidification (Croft 2003: 257-258) process following an earlier period in which both orders were possible in Latin (see Bauer 2009: 263f for diachronic analysis), we would be sorely disappointed to find that noun-adjective word order in Romance languages is in fact (still) so variable. A rigidification process needs to account for the contemporary variation, which implies that in the first place, we need to know about the variation.

Secondly, we may come back to implicational language universals. For example, the correlation between noun-genitive order and verb initial or VO order (Payne 1990: 14, Konstanz Archive universal #1549; Dryer 1986: 102, Konstanz Archive universal #1017) and between genitive-noun order and verb final or OV order (Lehmann 1973: 48, Konstanz Archive universal #107) is one of the strongest word order universals. If we took the WALS (Dryer & Haspelmath 2013) data (as done by Jäger *et al.* 2017 and Dunn *et al.* 2011), Dutch is counted as a noun-genitive order language, while it clearly has both orders (see Section 1). Knowing that such noise is present even for well-described, high-resource languages as Dutch, shows that categorical classifications potentially hide a lot of variation, aside from the fact that WALS allows for classifying languages with 'both orders' in a separate category. Measuring implies abstraction, but we feel that the amount of data reduction with regard to measurement in typology has been exceptional (Wälchli 2009). Conversely, implicational universals can and should be reformulated to take into account for such variation in sensible ways.

Third, in our implementation of a corpus-based typology we do not only apply quantitative methods in order to shed light on a given phenomenon, but we also try to illustrate some explanations for it. We do so intermittently for adnominal word order in Section 4.2, and there are two major explanations which we wish to return to shortly here: language change in process and information structure. A lot has been said on the importance and complexity of these factors (Croft 2003: 257-258 and Hawkins 1983: 213 on diachronic change in word order, and Givón 1988, Gundel 1988, Herring 1990, Payne 1990 on the interaction with information structure), and our analyses are in line with this body of work. Bauer (2009: 263f) describes variability in noun-adjective order in Romance as a consequence of earlier variation. However, as the discussion in Section 4.2 shows, this variability is rooted in a complex interplay of lexical constraints, pragmatics and stylistics. The same applies to Dutch *EDH-postpositions*, which are used in the context of previously established locations or times. Further explanations for picking one word order over the other, for the languages in which this is possible, may be rooted in cognitive explanations (Hawkins 1983, Hawkins 1990, Karimi, Diaz & Ferreira 2019) and these may impact different word orders to different extents (Levshina 2019, Östling & Wälchli 2018). These are valuable venues to explore further using more fine-grained and appropriate measurement, as proposed here. We believe that this discussion leaves us at an interesting junction, with tools falling into place that will allow us to investigate word order and other typological questions from a token-based perspective with an immediate eye on explanations for variability in terms of diachronic analysis and information management.

5. Conclusion

In this paper we have discussed an implementation of a corpus-based typology, presented a new parallel corpus and applied our methodology to a case-study. Our parallel corpus is called CIEP+, the Corpus of Indo-European Prose Plus, an ongoing prose collection of a balanced sample of Indo-European languages and non-Indo-European languages. We measure word order variability using entropy and use the models available in the Universal Dependencies project to generate morpho-syntactic annotation for eleven Indo-European languages in CIEP (closely related Germanic and Romance languages as well as Modern Greek, Polish, and Welsh).

The case-study showcased the striking difference between categorical classifications of word order and a token-based, corpus-based typology of word order of a set of five adnominal modifiers, here recast as comparative concepts: ANALYTIC CASE MARKER-NOUN, RELATIVE CLAUSE-NOUN, MODIFYING ADJECTIVE-NOUN, ARTICLE-NOUN and DEMONSTRATIVE-NOUN. We observed that nominal modifier word orders differ in their variability, both across modifiers and across languages. Low word order variability is generally attested for ANALYTIC CASE MARKER, ARTICLE, DEMONSTRATIVE and RELATIVE CLAUSE, and high variability is attested for MODIFYING ADJECTIVE. However, even for word orders with generally low variability, we find outliers: Dutch has somewhat variable ANALYTIC CASE MARKER-NOUN word order; Greek, Polish and Welsh have variable DEMONSTRATIVE-NOUN word order.

Our main contribution to the venture of token-based typology is providing a workable solution to make UD-based parsers more cross-linguistically applicable and showing potential register variation, both of which should be elaborated upon in future work.

Author contributions

LT: conceptualization; validation; data collecting; methodology; writing (original draft) sections: 2, 3, 4, Appendix; writing (review & editing).
AV: conceptualization; validation; data collecting; writing (original draft) sections: 1, 3, 4, 5; writing (review & editing).
All authors contributed to the article and approved the submitted version.

Abbreviations

1 = first person; 2 = second person; 3 = third person; ACC = accusative; ACM = analytic case marker; ADJ = adjective; ADP = adposition; ADV = adverb; AOR = aorist; CIEP+ = Corpus of Indo-European Prose Plus; COND = conditional; DAT = dative; DEF = definite; F = feminine; FORM = formal; FUT = future; INDF = indefinite; INF = infinitive; IPFV = imperfective; LOC = locative; LoL = Lists of Lemmata; N = neuter; NEG = negation/negative; NOM = nominative; PFV = perfective; PL = plural; PRF = perfect; PRON = pronoun; PROX = proximate; PRS = present; PST = past; PTCP = participle; REFL = reflexive; REL = relative; SG = singular; UD = Universal Dependencies project; UPOS = Universal Dependencies POS tagset.

Acknowledgments

This work is supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Project-ID 232722074 – SFB 1102.

Supplementary material

The supplementary material for this article can be found here: <doi.org/10.5281/zenodo.7082292>.
It contains:
- the Appendix: “An overview of CIEP+” (A), “The subset of CIEP used for the case-study” (B), “List of models” (C), “Comparative concepts and Universal Dependencies” (D), “Frequency values of the word order patterns (E)”;
- the R script used for the analysis;
- the Python script used for extracting word order patterns from CIEP;
- the dataset (head-dependent pairs and additional information).

Notes

[1] See <linguistic-typology.org/databases>.

[2] It may also be possible to use comparable units from original texts. According to Croft (2016: 378-379), token-based typology is a natural endpoint of using function or semantics as the basis of a comparative concept; since broad categorical classification is too coarse for cross-linguistic comparison, fine-grained information, down to the level of an experimental stimulus or a translation context, can provide a basis of comparison which is of the required granularity.

[3] Note that Dryer's threshold plays out differently in word orders where there are two logical orders, for example the order of adjective and noun (Dryer 2013b) or the order of adposition and noun (Dryer 2013c), as there are less possible orders. For such binary categorical variables as the order of adjective and noun, if adjective-noun is attested 67% of the time in a text (or more frequently), and noun-adjective order is attested 33% of the time (or less frequently), the language is classified as having dominant adjective-noun order. In these and similar WALS chapters, Dryer employs a third category in case a dominant order cannot be established, along the lines of "Both orders of noun and modifying adjective occur, with neither dominant" (Dryer 2013b), which implies that the least frequent order is attested in 34% of a text count or more frequently.

[4] Here we write *mijn leraars boek* instead of *leraars boek*, which would be in parallel with the other examples, because the latter sounds marginal at best, probably because *leraar* by itself is not specific enough to enable reference resolution to a single person, which is required for the use of the -*s*-genitive. It is fine when preceded with the possessive pronoun *mijn* 'mine', see also Weerman & De Wit (1999).

[5] <linguatools.org/tools/corpora/wikipedia-comparable-corpora>.

[6] Another benefit of entropy as a single measure of variability is the possibility to correlate variability in one domain with variability in another; for example, one can investigate the correlation between overt dependent- or head-marking with various dominant or free word orders (Dryer 2002; Futrell, Mahowald & Gibson 2015; Levshina 2019; Nichols 1986, 1992; Sinnemäki 2008, 2010) by looking at variability of dependent-marking, head-marking, and main clausal word order.

[7] A forerunner of UD is HamleDT (Zeman, Dušek *et al.* 2014), developed partly by the same team as UD. An alternative to UD is Surface Syntactic Universal Dependencies (SUD: Gerdes, Guillaume *et al.* 2018), which claims "a nearly perfect degree of two-way convertibility with the Universal Dependencies scheme" (<surfacesyntacticud.github.io>). UD-based approaches are predated by several other projects, which however cover only a handful of languages: for instance, the Parallel Grammar Project, which is active since the early nineties and works on cross-linguistic parsing using the Lexical-Functional Grammar framework (M. Butt *et al.* 2002).

[8] <universaldependencies.org/u/dep/index.html>.

[9] <universaldependencies.org/ext-dep-index.html>.

[10] We use the following typographic conventions. Universal Dependency (UD) Relations are written between single quotation marks, e.g. 'amod' for 'Adjectival modification', while Universal Parts of Speech (UPOS) are in uppercase format, e.g. ADJ for 'Adjectives' (Marneffe, Manning *et al.* 2021). Comparative concepts are written in small capitals, e.g. MODIFYING ADJECTIVE (Haspelmath 2010) and language-specific categories are written in regular font, e.g. English adjectives.

[11] <universaldependencies.org/format.html>.

[12] <www.uni-saarland.de/lehrstuhl/verkerk.html>.

[13] <ufal.mff.cuni.cz/udpipe/1/models>.

[14] Version 2.3.3: <github.com/pyconll/pyconll>.

[15] For this and other information, such as raw frequency data and scripts used to extract and analyse data, we refer to the Supplementary Material, which can be accessed here: <doi.org/10.5281/zenodo.7082292>.

[16] The German GSD treebank serving as the training data has actually six instances of 'acl:relcl' (<universaldependencies.org/treebanks/de_gsd/de_gsd-dep-acl-relcl.html>), but they are probably not enough to teach the parser to recognize relative clauses.

[17] See <en.wiktionary.org/wiki/Category:Polish_relational_adjectives> for a list of Polish relational adjectives.

[18] According to Maiden & Robustelli (2013: 171-172), in complex prepositions such

as *insieme a* 'together (with)' the first element can be stranded, as in *La persona alla quale vivo insieme* 'The person to whom I live with'. However, this syntactic relation is an instance of adverbial modification, with a verbal head and is recognized as such by the UD parser, which treats it as an 'advmod' (<universaldependencies.org/u/dep/advmod.html>).

[19] References for each word order are: adposition-noun order: Dryer (2013c); relative clause-noun order: Dryer (2013f); adjective-noun order: Dryer (2013b); demonstrative-noun order: Dryer (2013d); article-noun order: Tallerman (1998: 37) for Welsh; Holmberg & Rijkhoff (1998: 96) for Danish, Dutch, English, and German; Lascaratou (1998: 163-164) for Greek, and Arnaiz (1998: 62-63) for French, Italian, Portuguese, and Spanish.

[20] Current UD treebanks at <universaldependecies.org>. See Appendix C (Supplementary Material) for an overview of UD treebank sources for our sample.

[21] This last point is nicely illustrated by the two parallel examples discussed in Sections 2 and 4.2.3. By comparing the English and French translations of a lengthy sentence from Eco's *Il nome della ros*a, the first example shows the cross-linguistic variability of MODIFYING ADJECTIVE-NOUN order at the token level; the second parallel example provides instead a window on the intra-linguistic variability of this comparative concept, contrasting the translations in a sentence from Süskind's *Das Parfum* in four Romance languages.

Bibliographical References

Ackerman, Farrell & Malouf, Robert 2013. Morphological Organization: The Low Conditional Entropy Conjecture. *Language* 89,3. 429-464.

Altenberg, Bengt & Granger, Sylviane 2002. Recent Trends in Cross-Linguistic Lexical Studies. In Altenberg, Bengt & Granger, Sylviane (eds.), *Studies in Corpus Linguistics, Vol. 7*. Amsterdam: Benjamins. 3-48. <doi.org/10.1075/scl.7.04alt>.

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

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

Basirat, Ali; de Lhoneux, Miryam; Kulmizev, Artur; Kurfalı, Murathan; Nivre, Joakim & Ostling, Robert 2019. Polyglot Parsing for One Thousand and One Languages (And Then Some). *First workshop on Typology for Polyglot NLP*. Florence, Italy, 1 August 2019.

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

Bauer, Brigitte M. 2009. Word order. In Baldi, Philip & Cuzzolin, Pierluigi (eds.), *New Perspectives on Historical Latin Syntax. Vol. 1: Syntax of the Sentence.* Berlin: Mouton de Gruyter. 241-316.

Bentz, Christian; Alikaniotis, Dimitrios; Cysouw, Michael & Ferrer-i-Cancho, Ramon 2017. The Entropy of Words: Learnability and Expressivity across

More than 1000 Languages. *Entropy* 19,6. 275-307. <doi: 10. 3390/ e19060275>.

Bentz, Christian; Sozinova, Olga & Samardžić, Tanja 2019. Collecting a corpus for 100 typologically diverse languages (100LC). *Workshop on language documentation: Multilingual settings and technological advances*. 24-25 October 2019, Uppsala, 2019.

Berdicevskis, Aleksandrs; Çöltekin, Çağrı; Ehret, Katharina; von Prince, Kilu; Ross, Daniel; Thompson, Bill; Yan, Chunxiao; Demberg, Vera; Lupyan, Gary; Rama, Taraka & Bentz, Christian 2018. Using Universal Dependencies in cross-linguistic complexity research. *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*. Brussels, Belgium: Association for Computational Linguistics. 8-17. <doi: 10. 18653/v1/W18-6002>. <aclanthology.org/W18-6002>.

Berendsen, Bieneke 2021. EDH-postposition. <www.dutchgrammar.com/ en/?n=WordOrder.29>.

Biber, Douglas 1993. Using Register-Diversified Corpora for General Language Studies. *Computational Linguistics* 19,2. 219-241.

Biber, Douglas 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press.

Biber, Douglas 2012. Register as a Predictor of Linguistic Variation. *Corpus Linguistics and Linguistic Theory* 8,1. 9-37. <doi.org/10.1515/cllt-2012-0002>.

Bickel, Balthasar 2000. Referential Density in Discourse and Syntactic Typology. *Language* 79,4. 708-736. <doi.org/10.1353/lan.2003.0205>.

Bisang, Walter 2011. Word classes. In Song, Jae Jung (ed.), *Oxford Handbook of Linguistic Typology*. Oxford: Oxford University Press. 281-302.

Bizzoni, Yuri; Degaetano-Ortlieb, Stefania; Fankhauser, Peter & Teich, Elke 2020. Linguistic Variation and Change in 250 years of English Scientific Writing: A Data-driven Approach. *Frontiers in Artificial Intelligence.* <doi. org/10.3389/frai.2020.00073>.

Blake, Barry J. 2001. *Case*. Cambridge: Cambridge University Press.

Bochkarev, Vladimir; Solovyev, Valery D. & Wichmann, Søren 2014. Universals versus historical contingencies in lexical evolution. *Journal of the Royal Society Interface* 11. 1-8. <doi.org/10.1098/rsif.2014.0841>.

Bosco, Cristina; Dell'Orletta, Felice; Montemagni, Simonetta; Sanguinetti, Manuela & Simi, Maria. The Evalita 2014 Dependency Parsing task. *Proceedings of the Fourth International Workshop* EVALITA 2014. Pisa University Press. 1-8. <www.aclweb.org/anthology/W17-0413>.

Bouma, Gosse & van Noord, Gertjan 2017. Increasing Return on Annotation Investment: The Automatic Construction of a Universal Dependency Treebank for Dutch. *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017).* Gothenburg, Sweden: Association for Computational Linguistics. 19-26. <www.aclweb.org/anthology/W17-0403>.

Breindl, Eva 2020a. Definiter Artikel. *Institut für Deutsche Sprache: "Systematische Grammatik". Grammatisches Informationssystem grammis.* <doi: 10.14618/ grammatiksystem>.

Breindl, Eva 2020b. Demonstrativ-Pronomen. *Institut für Deutsche Sprache:*

"Systematische Grammatik". Grammatisches Informationssystem grammis. < doi: 10.14618/grammatiksystem > .

Breindl, Eva 2020c. Indefiniter Artikel. *Institut für Deutsche Sprache: "Systematische Grammatik". Grammatisches Informationssystem grammis*. < doi: 10.14618/grammatiksystem > .

Breindl, Eva 2020d. Präposition. *Institut für Deutsche Sprache: "Systematische Grammatik". Grammatisches Informationssystem grammis*. < doi: 10.14618/grammatiksystem > .

Breindl, Eva 2020e. Relativ-Elemente. *Institut für Deutsche Sprache: "Systematische Grammatik". Grammatisches Informationssystem grammis*. < doi: 10.14618/grammatiksystem > .

Broekhuis, Hans 2013. *Syntax of Dutch: Adpositions and Adpositional Phrases*. Amsterdam: Amsterdam University Press.

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

Butt, Miriam; Dyvik, Helge; Holloway King, Tracy; Masuichi, Hiroshi & Rohrer, Christian 2002. The Parallel Grammar Project. In Oostijk, N.; Carroll, J. & Sutcliffe, R. (eds.), *Proceedings of the Workshop on Grammar Engineering and Evaluation.* COLING02. 1-7.

Bybee, Joan L. 1988. The diachronic dimension in explanation. In Hawkins, John A. (ed.) *Explaining language universals.* Oxford: Blackwell. 350-379.

Cappelle, Bert 2012. English is less rich in manner-of-motion verbs when translated from French. *Across Languages and Cultures* 13,2. 173-195.

Chen, Xinying & Gerdes, Kim 2018. How Do Universal Dependencies Distinguish Language Groups? In Jian, Jingyang & Liu, Haitao (eds.), *Quantitative Analysis of Dependency Structures*. 277-294. isbn: 9783110573565. < doi: 10.1515/9783110573565014 > .

Cinque, Guglielmo 2010. *The Syntax of Adjectives. A Comparative Study. Cambridge*, Mass.: MIT Press.

Clackson, James 2007. *Indo-European Linguistics: An Introduction*. Cambridge: Cambridge University Press.

Cohen Priva, Uriel & Gleason, Emily 2016. Simpler structure for more informative words: A longitudinal study. *Proceedings of the 38th Annual Conference of the Cognitive Science Society.* Austin, TX: Cognitive Science Society. 1895-1900. < mindmodeling.org/cogsci2016/papers/0331/index.html > .

Collins, Jeremy 2019. Some language universals are historical accidents. In Schmidtke-Bode, Karsten; Levshina, Natalia; Michaelis, Susanne Maria & Seržant, Ilja (eds.), *Explanation in typology.* Berlin: Language Science Press. 47-61. < doi: 10.5281/zenodo.2583808 > .

Comrie, Bernard & Kuteva, Tania 2013. Relativization Strategies. *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. < wals.info/chapter/s8 > .

Coupé, Christophe *et al.* 2019. Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances* 5,9. < doi: 10.1126/sciadv.aaw2594 > . < advances.sciencemag.org/content/5/9/eaaw2594 > .

Croft, William 1990. A Conceptual Framework for Grammatical Categories (or: A

Taxonomy of Propositional Acts). *Journal of Semantics* 40,7. 245-279.

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

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

Croft, William 2016. Comparative concepts and language-specific categories: Theory and practice. *Linguistic Typology* 20,2. 377-393. <doi.org/10.1515/lingty-20160012>.

Croft, William *et al.* 2017. Linguistic Typology Meets Universal Dependencies. *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT15)*. Paris, France: CEUR Workshop Proceedings. 63-75. <ceur-ws.org/Vol-1779>.

Cysouw, Michael 2005. Quantitative Methods in Typology = Quantitative Methoden in der Typologie. In Altmann, Gabriel; Köhler, Reinhard & Piotrowski, Rajmund G. (eds.), *Quantitative Linguistik: Ein Internationales Handbuch.* Berlin: Walter de Gruyter. 554-578.

Degaetano-Ortlieb, Stefania & Teich, Elke 2019. Toward an optimal code for communication: The case of scientific English. *Corpus Linguistics and Linguistic Theory* (open access). 1-33. <www.degruyter.com/downloadpdf/j/cllt.ahead-ofprint/cllt-2018-0088/cllt-2018-0088.pdf>.

Dryer, Matthew S. 1986. Word order consistency and English. In DeLancey, Scott & Tomlin, Russel (eds.), *Proceedings of the Second Annual Pacific Linguistics Conference.* Eugene: University of Oregon. 97-106.

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

Dryer, Matthew S. 2002. Case distinctions, rich verb agreement, and word order type (Comments on Hawkins' paper). *Theoretical Linguistics* 28,2.

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

Dryer, Matthew S. 2013a. Determining Dominant Word Order. *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. <wals.info/chapter/s6>.

Dryer, Matthew S. 2013b. Order of Adjective and Noun. *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. <wals.info/chapter/87>.

Dryer, Matthew S. 2013c. Order of Adposition and Noun Phrase. *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <wals.info/chapter/85>.

Dryer, Matthew S. 2013d. Order of Demonstrative and Noun. *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <wals.info/chapter/88>.

Dryer, Matthew S. 2013e. Order of Genitive and Noun. *The World Atlas of Language Structures Online.* Leipzig: Max Planck Institute for Evolutionary Anthropology. <wals.info/chapter/86>.

Dryer, Matthew S. 2013f. Order of Relative Clause and Noun. *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <wals.info/chapter/90>.

Dryer, Matthew S. 2013g. Order of Subject, Object and Verb. *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <wals.info/chapter/81>.

Dryer, Matthew S. 2019. Grammaticalization accounts of word order correlations. In Schmidtke-Bode, Karsten; Levshina, Natalia; Michaelis, Susanne Maria & Seržant, Ilja (eds.), *Explanation in typology*. Berlin: Language Science Press. 63-95. <doi.org/10.5281/zenodo.2583810>.

Dryer, Matthew S. & Haspelmath, Martin (eds.) 2013. WALS Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. <wals.info>.

Dunn, Michael; Greenhill, Simon J.; Levinson, Stephen C. & Gray, Russell D. 2011. Evolved Structure of Language Shows Lineage-Specific Trends in Word-Order Universals. *Nature* 473. 79-82. <doi.org/10.1038/nature09923>.

Eisenberg, Peter & Schöneich, Rolf 2020. *Grundriss der deutschen Grammatik: Der Satz*. 5th ed. Stuttgart: J.B. Metzler.

Futrell, Richard; Levy, Roger P. & Gibson, Edward 2020. Dependency Locality as an Explanatory Principle for Word Order. *Language* 96,2. 371-412. <doi.org/10.1353/lan.2020.0024>.

Futrell, Richard; Mahowald, Kyle & Gibson, Edward 2015. Quantifying Word Order Freedom in Dependency Corpora. *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*. Uppsala, Sweden: Uppsala University. 91-100. <www.aclweb.org/anthology/W15-2112>.

Geertzen, Jeroen; Blevins, James P. & Milin, Petar 2016. The informativeness of linguistic unit boundaries. *Italian Journal of Linguistics* 28,1. 25-48.

Gellerstam, Martin 1996. Translations as a Source for Cross-Linguistic Studies. In Ajmer, K.; Altenberg, Bengt & Johansson, M. (eds.), *Languages in Contrast. Papers from a Symposium on Text-Based Cross-Linguistic Studies.* Lund: Lund University Press.

Gerdes, Kim; Guillaume, Bruno; Kahane, Sylvain & Perrier, Guy 2018. SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. *Universal Dependencies Workshop 2018.* Brussels, Belgium. <hal.inria.fr/hal-01930614>.

Gerdes, Kim; Kahane, Sylvain & Chen, Xinying 2019. Rediscovering Greenberg's Word Order Universals in UD. *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>. <www.aclweb.org/anthology/W198015>.

Gibson, Edward; Futrell, Richard; Piandadosi, Steven T.; Dautriche, Isabelle; Mahowald, Kyle; Bergen, Leon & Levy, Roger 2019. How Efficiency Shapes Human Language. *Trends in Cognitive Sciences* 23,5. 389-407.

Givón, Talmy 1988. The Pragmatics of Word Order: Predictability, Importance and Attention. In Hammond, Michael; Moravcsik, Edith A. & With, Jessica (eds.), *Studies in Syntactic Typology.* Amsterdam: Benjamins. 244-284.

Goldhahn, Dirk; Eckart, Thomas & Quasthoff, Uwe 2012. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC'12). Istanbul, Turkey: European

Language Resources Association (ELRA). 31-43.

Greenberg, Joseph H. 1960. A Quantitative Approach to the Morphological Typology of Language. *International Journal of American Linguistics* 26,3. 178-194.

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

Guillaume, Bruno; de Marneffe, Marie-Catherine & Perrier, Guy 2019. Conversion et améliorations de corpus du français annotés en Universal Dependencies. *Traitement Automatique des Langues* 60,2. 71-95. <hal.inria.fr/hal-02267418>.

Gundel, Jeanette 1988. Universals of Topic-Comment Structure. In Hammond, Michael; Moravcsik, Edith A. & With, Jessica (eds.), *Studies in Syntactic Typology.* Amsterdam: Benjamins. 209-239.

Hagège, Claude 2010. *Adpositions*. Oxford: Oxford University Press.

Hahn, Michael; Degen, Judith & Futrell, Richard 2021. Modeling word and morpheme order in natural language as an efficient trade-off of memory and surprisal. *Psychological Review.* 128,4. 726-756. <doi.org/10.1037/rev0000269>.

Haig, Geoffrey & Schnell, Stephan (eds.) 2021. Multi-CAST: Multilingual corpus of annotated spoken texts. <multicast.aspra.uni-bamberg.de>.

Hale, John 2016. Information-theoretical Complexity Metrics. *Language and Linguistics Compass* 10,9. 397-412. <doi.org/10.1111/lnc3.12196>.

Haspelmath, Martin 2001. The European Linguistic Area: Standard Average European. In Oesterreicher, Wulf; Haspelmath, Martin & Raible, Wolfgang (eds.), *Language Typology and Language Universals, Handbücher zur Sprach- und Kommunikationswissenschaft.* Berlin: Mouton de Gruyter. 1492-1510.

Haspelmath, Martin 2010. Comparative concepts and descriptive categories in cross-linguistic studies. *Language* 86,4. 663-687

Haspelmath, Martin 2018. How comparative concepts and descriptive linguistic categories are different. In Van Olmen, D.; Mortelmans, T. & Brisard, F. (eds.), *Aspects of Linguistic Variation.* Berlin: De Gruyter. 83-114.

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

Hawkins, John A. 1990. A Parsing Theory of Word Order Universals. *Linguistic Inquiry* 21,2. 223-261.

Hawkins, John A. 1994. *A Performance Theory of Order and Constituency*. Vol. 73. Cambridge Studies in Linguistics. Cambridge: Cambridge University Press.

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

Herring, Susan C. 1990. Information Structure as a Consequence of Word Order Type. *Annual Meeting of the Berkeley Linguistics Society* 16. 163-174.

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

Holmberg, Andres & Rijkhoff, Jan 1998. Word order in the Germanic languages.

Constituent Order in the Language of Europe. Berlin: Mouton de Gruyter. 75-104.

Holton, David; Mackridge, Peter & Philippaki-Warburton, Irene 2012. *Greek: A comprehensive grammar*. 2nd ed. London: Routledge.

Huddleston, Rodney & Pullum, Geoffrey K. 2002. *The Cambridge Grammar of the English language*. Cambridge: Cambridge University Press.

Jäger, Gerhard *et al.* 2017. The Evolution of Word-Order Universals: Some Word-Order Correlation Are Lineage Specific – Others Might Be Universal. Talk presented at the 12th Conference of the Association for Linguistic Typology, Canberra, December 12.

Johannsen, Anders; Martínez Alonso; Héctor & Plank, Barbara 2015. Universal Dependencies for Danish. *Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories* (TLT14). Warsaw, Poland: Institute of Computer Science, Polish Academy of Sciences. 157-167. <tlt14.ipipan.waw.pl/proceedings>.

Kalimeri, Maria; Constantoudis, Vassilios; Papadimitriou, Constantinos; Karamanos, Konstantinos; Diakonos, Fotis K. & Papageorgiou, Harris 2015. Word-length Entropies and Correlations of Natural Language Written Texts. *Journal of Quantitative Linguistics* 22,2. 101-118. <doi: 10.1080/09296174.2014.1001636>.

Kann, Katharina; Bowman, Samuel R. & Cho, Kyunghyun 2020. Learning to Learn Morphological Inflection for Resource-Poor Languages. *Proceedings of the AAAI Conference on Artificial Intelligence* 34,5. AAAI Press, Palo Alto, California USA. 8058-8065.

Karimi, Hossein; Diaz, Michele & Ferreira, Fernanda 2019. A Cruel King Is Not the Same as a King Who Is Cruel: Modifier Position Affects How Words Are Encoded and Retrieved from Memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 45,11. 2010-2035. <doi.org/10.1037/xlm0000694>.

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

Koehn, Philipp 2005. Europarl: A parallel corpus for statistical machine translation. *Proceedings of Machine Translation Summit X: Papers*. Phuket, Thailand. 79-86.

Koplenig, Alexander; Meyer, Peter; Wolfer, Sascha & Müller-Spitze, 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. <doi.org/10.1371/journal.pone.0173614>.

Kurzová, Helena 1981. *Der Relativsatz in den indoeuropäischen Sprachen*. Hamburg: Buske.

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

Lehmann, Winfred P. 1973. A Structural Principle of Language and Its Implications. *Language* 49,1. 47-66. <doi.org/10.2307/412102>.

Levshina, Natalia 2015. European analytic causatives as a comparative concept: Evidence from a parallel corpus of film subtitles. *Folia Linguistica* 49,2. 487-520. <doi.org/10.1515/ling-2012-0021>.

Levshina, Natalia 2017. Film Subtitles as a Corpus: An N-Gram Approach. *Corpora* 12,3. 311-338. <doi.org/10.1515/ling-2012-0021>.

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. Corpus-based typology: Applications, challenges and some solutions. *Linguistic Typology*. <doi.org/10.1515/lingty-2020-0118>.

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

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

Maiden, Martin & Robustelli, Cecilia 2013. *A reference grammar of modern Italian*. 2nd ed. London / New York: Routledge.

Majid, Asifa; Boster, James S. & Bowerman, Melissa 2008. The cross-linguistic categorization of everyday events: A study of cutting and breaking. *Cognition* 109,2. 235-250. <doi: 10.1016/j.cognition.2008.08.009>. <dx.doi.org/10.1016/j.cognition.2008.08.009>.

Majid, Asifa; Enfield, Nicholas J. & Van Staden, Myriam (eds.) 2006. Parts of the body: Crosslinguistic categorisation. *Language Sciences* 28,2-3. 137-360.

Majid, Asifa *et al.* 2018. Differential coding of perception in the world's languages. *Proceedings of the National Academy of Sciences* 115,45. 11369-11376. <doi.org/10.1073/pnas.1720419115>.

Marneffe, Marie-Catherine de; Dozat, Timothy; Silveira, Natalia; Haverinen, Katri; Ginter, Filip; Nivre, Joakim & Manning, Christopher D. 2014. Universal Stanford dependencies: A cross-linguistic typology. *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC'14). Reykjavik, Iceland: European Language Resources Association (ELRA). 4585-4592. <www.lrec-conf.org/proceedings/ lrec2014/pdf/1062_Paper.pdf>.

Marneffe, Marie-Catherine de; Manning, Christopher D.; Nivre, Joakim & Zeman, Daniel 2021. Universal Dependencies. *Computational Linguistics* 47,2. 255-308. <doi.org/10.1162/coli_a_00402>.

Mayer, Thomas & Cysouw, Michael 2014. Creating a massively parallel Bible corpus. *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC'14). Reykjavik, Iceland: European Language Resources Association (ELRA). 3158-3163. <www.lrec-conf.org/proceedings/lrec2014/pdf/220_Paper.pdf>.

McDonald, Ryan *et al.* 2013. Universal Dependency Annotation for Multilingual Parsing. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers). Sofia, Bulgaria: Association for Computational Linguistics. 92-97. <www.aclweb.org/ anthology/P13-2017>.

Miestamo, Matti 2013. Symmetric and Asymmetric Standard Negation. *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <wals.info/chapter/113>.

Montemurro, Marcelo A. & Zanette, Damián H. 2011. Universal Entropy of

Word Ordering Across Linguistic Families. *PLOS ONE* 6,5. 1-9. <doi.org/10.1371/journal.pone.0019875>.

Naranjo, Matias-Guzmán & Becker, Laura 2018. Quantitative word order typology with UD. *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories* (TLT 2018). Oslo, Norway. Linköping Electronic Conference Proceedings 155,10. 91-104.

Nichols, Johanna 1986. Head-Marking and Dependent-Marking Grammar. *Language* 62,1. 56-119.

Nichols, Johanna 1992. Linguistic diversity in space and time. Chicago: University of Chicago Press.

Nichols, Johanna & Bickel, Balthasar 2013. Locus of Marking in the Clause. *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <wals.info/ chapter/23>.

Östling, Robert & Wälchli, Bernhard 2018. Word order typology extracted from parallel texts, its evaluation, and its potential for exploring word order variability. Paper presented at the 51st Annual Meeting of the *Societas Linguistica Europaea.*

Payne, Doris L. 1990. *The Pragmatics of Word Order: Typological Dimensions of Verb Initial Languages.* Berlin: Mouton de Gruyter.

Pederson, Eric; Danziger, Eve; Wilkins, David; Levinson, Stephen; Kita, Sotaro & Senft, Gunther 1998. Semantic typology and spatial conceptualization. *Language* 74,3. 557-589.

Petrov, Slav; Das, Dipanjan & McDonald, Ryan 2012. A Universal Part-of-Speech Tagset. *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC'12). Istanbul, Turkey: European Language Resources Association (ELRA). 2089-2096. <www.lrec-conf.org/proceedings/lrec2012/pdf/274_Paper.pdf>.

Prokopidis, Prokopis & Papageorgiou, Haris 2017. Universal Dependencies for Greek. *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017).* Gothenburg, Sweden: Association for Computational Linguistics. 102-106. <www.aclweb.org/anthology/W17-0413>.

San Roque, Lila *et al.* 2015. Vision verbs dominate in conversation across cultures, but the ranking of non-visual verbs varies. *Cognitive Linguistic*s 26,1. 1-30. <doi: 10.1515/ cog-2014-0089>. <www.degruyter.com/view/j/cogl.2015.26.issue1/cog-2014-0089/cog-2014-0089.xml>.

Santos, Diana 1995. On grammatical translationese. *Short papers presented at the Tenth Scandinavian Conference on Computational Linguistic*s. 59-66.

Seifart, Frank; Strunk, Jan; Danielsen, Swintha; Hartmann, Iren; Pakendorf, Brigitte; Wichmann, Søren; Witzlack-Makarevich, Alena; de Jong, Nivja H. & Bickel, Balthasar 2018. Nouns slow down speech across structurally and culturally diverse languages. *Proceedings of the National Academy of Sciences* 115,22. 5720-5725.

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

Siewierska, Anna & Uhlířová, Ludmilla 1998. Word order in the Slavic languages. *Constituent Order in the Language of Europe*. Berlin: Mouton de Gruyter. 105-149.

Sinnemäki, Kaius 2008. Complexity trade-offs in core argument marking.

Miestamo, Matti; Sinnemäki, Kaius & Karlsson, Fred (eds.), *Language complexity: Typology, contact, change*. Amsterdam: Benjamins. 67-88.

Sinnemäki, Kaius 2010. Word order in zero-marking languages. *Studies in Language* 34,4. 869-912. <doi.org/10.1075/sl.34.4.04sin>.

Slobin, Dan I.; Bowermann, Melissa; Brown, Penelope; Eisenbeiß, Sonja & Narasimhan, Bhuvana 2011. Putting Things in Places: Developmental Consequences of Linguistic Typology. In Bohnemeyer, Jürgen & Pederson, E. (eds.), *Event representation*. Cambridge: Cambridge University Press. 134-165.

Sorace, Antonella 2000. Gradients in Auxiliary Selection with Intransitive Verbs. *Language* 76,4. 859-890.

Stassen, Leon 2013. Noun Phrase Conjunction. *The World Atlas of Language Structures Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. <wals.info/chapter/63>.

Stivers, Tanya *et al.* 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences* 106,26. 10587-10592. <doi.org/10.1073/pnas.0903616106>.

Stolz, Thomas & Gugeler, Traude 2000. Comitative Typology. *STUF-Sprachtypologie Und Universalienforschung* 53,1. 53-61.

Stolz, Thomas; Stroh, Cornelia & Aina, Urdze 2006. *On comitatives and related categories: A typological study with special focus on the languages of Europe.* Berlin: Mouton de Gruyter.

Straka, Milan & Straková, Jana 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Vancouver, Canada: Association for Computational Linguistics. 88-99. <www.aclweb.org/anthology/K/K17/K173009.pdf>.

Swan, Oscar E. 2002. *A grammar of contemporary Polish*. Bloomington: Slavica.

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

Wälchli, Bernhard 2007. Advantages and Disadvantages of Using Parallel Texts in Typological Investigations. *STUF-Sprachtypologie Und Universalienforschung* 60,2. 118-134.

Wälchli, Bernhard 2009. Data Reduction Typology and the Bimodal Distribution Bias. *Linguistic Typology* 13,1. 77-94.

Wälchli, Bernhard 2019. The Feminine Gender Gram, Incipient Gender Marking, Maturity, and Extracting Anaphoric Gender Markers from Parallel Texts. In Olsson, Bruno; Di Garbo, Francesca & Wälchli, Bernhard (eds.), *Grammatical gender and linguistic complexity II: World-wide comparative studies*. Berlin: Language Science Press. 61-131.

Wälchli, Bernhard & Cysouw, Michael 2012. Lexical Typology through Similarity Semantics: Toward a Semantic Map of Motion Verbs. *Linguistics* 50,3. 118-134. <doi.org/10.1515/ling-2012-0021>.

Waldenfels, Ruprecht von 2006. Compiling a parallel corpus of Slavic languages. Text strategies, tools and the question of lemmatization in alignment. In Brehmer, B.; Zdanova, V. & Zimny, R. (eds.), *Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV)* 9. München: Otto Sagner. 123-138.

Weerman, Fred & De Wit, Petra 1999. The Decline of the Genitive in Dutch. *Linguistics* 37,6. 1155-1192. <doi.org/10.1515/ling.37.6.1155>.

Whitlam, John 2011. Modern Brazilian Portuguese Grammar: A Practical Guide. London / New York: Routledge.

Williams, Stephen J. 1980. A Welsh grammar. Cardiff: University of Wales Press.

Wróblewska, Alina 2018. Extended and Enhanced Polish Dependency Bank in Universal Dependencies Format. In de Marneffe, Marie-Catherine; Lynn, Teresa & Schuster, Sebastian (eds.), *Proceedings of the Second Workshop on Universal Dependencies* (UDW 2018). Association for Computational Linguistics. 173-182.

Zeldes, Amir 2017. The GUM Corpus: Creating Multilayer Resources in the Classroom. *Language Resources and Evaluation* 51,3. 581-612. <doi.org/10.1007/s10579-016-9343-x>.

Zeman, Daniel 2008. Reusable Tagset Conversion Using Tagset Drivers. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA). <www.lrec-conf.org/proceedings/lrec2008/pdf/66_paper.pdf>.

Zeman, Daniel; Dušek, Ondřej; Mareček, David; Popel, Martin; Ramasamy, Loganathan; Štěpánek, Jan; Žabokrtský, Zdeněk & Hajič, Jan 2014. HamleDT: Harmonized Multi-Language Dependency Treebank. *Language Resource and Evaluation* 48,4. 601-637. <doi: 10. 1007/s10579-014-9275-2>.

Zeman, Daniel *et al.* 2020. Universal Dependencies 2.6. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. <hdl.handle.net/11234/1-3226>.