# Supplementary material accompanying “A new methodology for an old problem: a corpus-based typology of adnominal word order in European languages” by Luigi Talamo and Annemarie Verkerk – Appendix

# Appendix A: An overview of CIEP+

Initially started as the data source for a PhD project and featuring just a handful of novels in thirty languages, the Corpus of Indo-European Prose Plus is currently further in development and will include 33 Indo-European languages and 10 languages from other families. In the table below we list the languages featured in CIEP+, together with the number of books and the entropy of four adnominal word order patterns; the entropy values come from Levshina (2019), which uses UD 2.1 treebanks as data source, while the four adnominal word order patterns correspond to the five comparative concepts investigated in our case-study.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Language** | **Family** | **(sub)family** | **books** | **adp\_Noun** | **amod\_Noun** | **det\_Noun** | **acl\_Noun** |
| Albanian | IE | Albanian | 17 | - | - | - | - |
| Armenian | IE | Armenian | 14 | - | - | - | - |
| Bulgarian | IE | Balto-Slavic | 18 | 0.000 | 0.199 | 0.930 | 0.051 |
| Croatian-Serbian | IE | Balto-Slavic | 18 | 0.010 | 0.081 | 0.024 | 0.045 |
| Czech | IE | Balto-Slavic | 18 | 0.003 | 0.367 | 0.121 | 0.018 |
| Latvian | IE | Balto-Slavic | 17 | 0.094 | 0.024 | 0.097 | 0.119 |
| Lithuanian | IE | Balto-Slavic | 18 | 0.074 | 0.101 | 0.000 | 0.761 |
| Polish | IE | Balto-Slavic | 18 | 0.058 | 0.931 | 0.317 | 0.000 |
| Russian | IE | Balto-Slavic | 18 | 0.018 | 0.114 | 0.154 | 0.703 |
| Breton | IE | Celtic | 4 | - | - | - | - |
| Irish | IE | Celtic | 5 | 0.013 | 0.100 | 0.688 | 0.000 |
| Welsh | IE | Celtic | 5 | - | - | - | - |
| Danish | IE | Germanic | 18 | 0.077 | 0.255 | 0.011 | 0.053 |
| Dutch | IE | Germanic | 18 | 0.035 | 0.168 | 0.006 | 0.459 |
| English | IE | Germanic | 18 | 0.163 | 0.233 | 0.001 | 0.227 |
| German | IE | Germanic | 18 | 0.061 | 0.026 | 0.009 | 0.158 |
| Swedish | IE | Germanic | 17 | 0.049 | 0.104 | 0.003 | 0.020 |
| Modern Greek | IE | Hellenic | 18 | 0.016 | 0.177 | 0.131 | 0.295 |
| Bengali | IE | Indo-Aryan | 13 | - | - | - | - |
| Gujarati | IE | Indo-Aryan | 4 | - | - | - | - |
| Hindi | IE | Indo-Aryan | 16 | 0.087 | 0.014 | 0.005 | 0.988 |
| Marathi | IE | Indo-Aryan | 13 | 0.000 | 0.000 | 0.000 | NA |
| Nepali | IE | Indo-Aryan | 4 | - | - | - | - |
| Punjabi | IE | Indo-Aryan | 4 | - | - | - | - |
| Sinhala | IE | Indo-Aryan | 12 | - | - | - | - |
| Urdu | IE | Indo-Aryan | 8 | 0.180 | 0.057 | 0.000 | 0.795 |
| Persian | IE | Iranian | 16 | 0.548 | 0.351 | 0.000 | 0.000 |
| French | IE | Romance | 18 | 0.002 | 0.872 | 0.002 | 0.183 |
| Italian | IE | Romance | 18 | 0.001 | 0.905 | 0.020 | 0.101 |
| Latin | IE | Romance | 6 | 0.006 | 0.998 | 0.871 | 0.386 |
| Portuguese | IE | Romance | 17 | 0.009 | 0.876 | 0.036 | 0.149 |
| Romanian | IE | Romance | 18 | 0.014 | 0.530 | 0.561 | 0.042 |
| Spanish | IE | Romance | 18 | 0.035 | 0.830 | 0.019 | 0.048 |
| Arabic | Afroasiatic | Semitic | 18 | 0.018 | 0.088 | 0.416 | 0.011 |
| Basque | isolate | isolate | 14 | 0.000 | 0.849 | 0.942 | 0.153 |
| Finnish | Uralic | Finnic | 17 | 0.614 | 0.038 | 0.022 | 0.986 |
| Georgian | Kartvelian | Georgian | 17 | - | - | - | - |
| Hungarian | Uralic | Hungarian | 18 | 0.000 | 0.008 | 0.000 | 0.000 |
| Indonesian | Austronesian | Malayo-Polynesian | 16 | 0.122 | 0.397 | 0.994 | 0.040 |
| Japanese | Japonic | Japanesic | 17 | 0.000 | 0.000 | 0.000 | 0.000 |
| Mandarin Chinese | Sino-Tibetan | Sinitic | 18 | 0.988 | 0.022 | 0.003 | 0.952 |
| Tamil | Dravidian | S. Dravidian | 8 | 0.000 | 0.019 | 0.082 | 0.619 |
| Turkish | Turkic | Turkish | 18 | 0.085 | 0.052 | 0.014 | 0.028 |

CIEP language sample with details on genealogy, size of the subcorpus and entropy of adnominal word order. Information on adnominal word order is taken from Levshina (2019) (Supplementary Material: entropy60\_languages\_24\_patterns.xls).

As the name suggests, CIEP+ is a collection of literary works and their translations; the following is a list with full references to the original books featured in CIEP; each book is identified with a short name in small capitals.

1. AAiW: L. Carroll. *Alice’s Adventures in Wonderland.* London: Macmillan, 1865.
2. Achterhuis: A. Frank, O. Frank, M. Pressler. *Het Achterhuis: dagboekbrieven 12 juni 1942 - 1 augustus 1944*. Uitgeverij Bert Bakker, 1947.
3. Alquimista: P. Coelho. *O Alquimista.* Rio de Janeiro: Rocco, 1989.
4. CienAños: G. García Márquez. *Cien Años de Soledad*. Buenos Aires: Editorial Sudamericana , 1967.
5. JFeN: G. Musso. *La jeune fille et la nuit.* Paris: Calmann-Lévy, 2018.
6. NomeRosa: U. Eco. *Il nome della rosa* Milano: Bompiani, 1980.
7. Parfum: P. Süskind. *Das Parfum: Die Geschichte eines Mörders*. Zürich: Diogenes, 1985.
8. PetitPrince: A. de Saint-Exupery. *Le Petit Prince*. Paris: Gallimard, 1943.
9. Potter1: J.K Rowling. *Harry Potter and the Philosopher’s Stone*. London: Bloomsbury Publishing Plc., 1997.
10. Potter2: J.K Rowling. *Harry Potter and the Chamber of Secrets*. London: Bloomsbury Publishing Plc., 1998.
11. Potter3: J.K Rowling. *Harry Potter and the Prisoner of Azkaban*. London: Bloomsbury Publishing Plc., 1999.
12. Potter4: J.K Rowling. *Harry Potter and the Goblet of Fire.* London: Bloomsbury Publishing Plc., 2000.
13. Potter5: J.K Rowling. *Harry Potter and the Order of the Phoenix*. London: Bloomsbury Publishing Plc., 2003.
14. Potter6: J.K Rowling. *Harry Potter and the Half-Blood Prince*. London: Bloomsbury Publishing Plc., 2005.
15. Potter7: J.K Rowling. *Harry Potter and the Deathly Hallows.* London: Bloomsbury Publishing Plc., 2007.
16. TtLG: L. Carroll. *Through the Looking-Glass and What Alice Found There.* London: Macmillan, 1871.
17. Zahir: P. Coelho. *O Zahir*. Rio de Janeiro: Rocco, 2005.
18. Zorba: N. Kazantzakis. Βίος και Πολιτεία του Αλέξη Ζορμπά (Víos kai Politeía tou Aléxē Zorbá). Athens: Dēmētriou Dēmētrakou A.E., 1946.

# Appendix B: The subset of CIEP used for the case-study

The table below gives for each book the year of translation into the eleven languages of the sample used for the case-study; languages are identified by their language code; the publishing year of the original book is given in bold; a hyphen ‘-’ indicates that the corresponding translation does not exist.

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **da** | **nl** | **en** | **fr** | **de** | **el** | **it** | **pl** | **pt** | **es** | **cy** |
| AAiW | 1946 | 2006 | **1865** | 2009 | 1999 | 2003 | 1914 | 2010 | 2000 | 1986 | 1982 |
| Achterhuis | 2006 | **1947** | 1995 | 1987 | 1955 | 1990 | 2003 | 2020 | 1996 | 1957 | 1986 |
| Alquimista | 2015 | 1994 | 2006 | 1995 | 1996 | 1996 | 1995 | 1995 | **1989** | 2002 | - |
| CienAños | 2014 | 1972 | 1970 | 1968 | 1984 | 1982 | 2005 | 2017 | 1995 | **1967** | - |
| JFeN | 2019 | 2019 | 2019 | **2018** | 2019 | 2019 | 2018 | 2019 | - | 2018 | - |
| NomeRosa | 2013 | 1985 | 2011 | 1982 | 1986 | 2011 | **1980** | 1980 | 2011 | 2015 | - |
| Parfum | 2005 | 1997 | 2006 | 1985 | **1985** | 2016 | 1988 | 1990 | 1985 | 1985 | - |
| PetitPrince | 1962 | 1980 | 2000 | **1943** | 2015 | 2007 | 1980 | 1985 | 1960 | 1951 | 2007 |
| Potter1 | 1997 | 1998 | **1997** | 1998 | 1998 | 1998 | 1998 | 2000 | 1999 | 1998 | 2003 |
| Potter2 | 1999 | 1999 | **1998** | 1999 | 1999 | 1999 | 1999 | 2000 | 2000 | 1999 | - |
| Potter3 | 1999 | 2000 | **1999** | 1999 | 1999 | 1999 | 2000 | 2001 | 2000 | 2000 | - |
| Potter4 | 2000 | 2000 | **2000** | 2000 | 2000 | 2000 | 2001 | 2001 | 2001 | 2001 | - |
| Potter5 | 2003 | 2003 | **2003** | 2003 | 2003 | 2003 | 2003 | 2004 | 2003 | 2004 | - |
| Potter6 | 2005 | 2005 | **2005** | 2005 | 2005 | 2005 | 2006 | 2006 | 2005 | 2006 | - |
| Potter7 | 2007 | 2007 | **2007** | 2007 | 2007 | 2007 | 2008 | 2008 | 2007 | 2008 | - |
| TtLG | 1946 | 2006 | **1871** | 2009 | 2000 | 1979 | 1914 | 2010 | 2000 | 1972 | 1984 |
| Zahir | 2015 | 2005 | 2005 | 2005 | 2005 | 2005 | 2017 | 2005 | **2005** | 2005 | - |
| Zorba | 1954 | 2015 | 2014 | 1997 | 2002 | **1946** | 1992 | 1996 | 1983 | 1988 | - |

# Appendix C: List of models

The following table lists the models used for parsing the texts, together with their references. For all languages except Welsh we have used the pre-trained UD 2.5 models available at <https://ufal.mff.cuni.cz/udpipe/1/models>; since a pre-trained model was not available for Welsh, we have trained our model using the training file from UD Welsh CCG 2.5.

|  |  |  |  |
| --- | --- | --- | --- |
| **Language** | **Model** | **Genre** | **Reference** |
| Danish (da) | UD Danish DDT v2.5 | fiction, news, non-fiction, spoken | Johannsen et al. (2015) |
| Dutch (nl) | UD Dutch Alpino v2.5 | news | Bouma and Noord (2017) |
| English (en) | UD English GUM v2.5 | academic, fiction, news, non-fiction, spoken, web, wiki | Zeldes (2017) |
| French (fr) | UD French GSD v2.5 | blog, news, reviews, wiki | Guillaume, Marneffe, and Perrier (2019) |
| German (de) | UD German GSD v2.5 | news, reviews, wiki | McDonald et al. (2013) |
| Greek (el) | UD Greek GDT v2.5 | news, spoken, wiki | Prokopidis and Papageorgiou (2017) |
| Italian (it) | UD Italian ISDT v2.5 | legal, news, wiki | Bosco et al. (2014) |
| Polish (pl) | UD Polish PDB v.2.5 | fiction, news, non-fiction | Wróblewska (2018) |
| Portuguese (pt) | UD Portuguese GSD v.2.5 | blog, news | McDonald et al. (2013) |
| Spanish (es) | UD Spanish GSD v.2.5 | blog, news, reviews, wiki | McDonald et al. (2013) |
| Welsh (cy) | UD Welsh CCG v.2.5 | fiction, grammar-examples, news, non-fiction, wiki | Heinecke and Tyers (2019) |

# Appendix D: Comparative concepts and Universal Dependencies

## Noun

Among the different proposals on defining the lexical category 'noun' (see Bisang 2011 for an overview), we find the one rooted in Radical Constructional Grammar particularly convincing.

According to Croft (2001: 88), lexical categories can be defined as constructions combining propositional (discourse) acts (Croft 1990: 247–248), such as reference, predication and modification, with semantic categories[[1]](#footnote-1), such as object, actions and property concepts.

A noun22 is thus a construction combining the discourse act of reference and the semantic category of object, or, put in shorter terms:

A noun makes a reference to an object.

The syntactic relation between the noun and the five modifiers is already specified at the modifier level, so its implementation does not require additional UD Relations; the implementation of the comparative concept can be made more accurate by using the NOUN (common nouns) and PROPN (proper nouns) from the UPOS tagset.

## Relative clause

The following comparative concept:

Within a noun phrase, a relative clause is a clause that is used to narrow the reference of the nominal head.

echoes the definition given in the WALS chapter devoted to relativization strategies (Comrie and Kuteva 2013), combining it with the comparative concept of adjective given by Haspelmath (2010: 670).

The similarity of relative clauses and adjectives is also highlighted by the UD classification of relative clauses as a type of adjectival clauses, or 'acl', which includes all types of clausal strategies modifying a noun phrase. The notion of modification is taken quite broad in UD and the 'acl' relation includes strategies such as ‘optional depictives’ and ‘finite clausal complements for nouns with a subset of nouns like fact or report’[[2]](#footnote-2).

In order to distinguish relative clauses from other adjectival clauses, some UD language models employ a dedicated relation subtype, 'acl:relcl'[[3]](#footnote-3); in our language sample, this relation subtype is used in all language models with the exception of the German model. In principle, it would be possible to recognize relative clauses even in language models missing the 'acl:relcl' relation by narrowing the comparative concept to reflect further characteristics. Accordingly, the proposed comparative concept could be re-formulated as follows, in order to account for relative clause in German and many other Standard Average European languages (Haspelmath 2001: 1494–1495):

Within a noun phrase, a relative clause is a clause that is used to narrow the reference of the nominal head and is introduced by a pronoun, which indexes through inflection the nominal head’s role within the relative clause.

The inflecting relative pronoun can be recognized through a language-specific list of *lemmata*, while the indexical relationship between the pronoun and the nominal head as well as its role within the relative clause is taken care by the UD Relations; however, the relative clause as re-formulated above corresponds to a particular and cross-linguistically unique type of relativizing strategy. Moreover, the re-formulated comparative concept excludes one of the languages of our sample, Welsh, which uses a relative particle to introduce the relative clause; this type of relativizing strategy is also very frequent in the other languages of our sample.

In order to capture the whole range of relativizing strategies, we use here a broader comparative concept for relative clause. We employ the more specific 'acl:relcl' relation where possible, but 'acl' for German, thus implying that data from German will also include other types of adjectival clauses.

The 'acl:relcl' and 'acl' relations are found on the head of the relative clause, which consists of verbs (UPOS: VERB) for verbal predicates, and of adjectives and nouns (UPOS: NOUN/PROPN and ADJ) for non-verbal predicates.

## Modifying Adjective

Since the typological research on the relative position of the adjective and noun has focused on modifying adjectives (‘descriptive adjectives’ in Greenberg 1963, see remarks in Hawkins 1983, Dryer 2013b), we propose the following comparative concept based on Haspelmath (2010: 670):

Within a noun phrase, a modifying adjective is a lexeme that denotes a descriptive property and is used to narrow the reference of the nominal head through its modification.

Our definition makes use of the ‘modification’ function; along with the predication and the reference function, which we have discussed at the beginning of this section when introducing the noun comparative concept, modification is one of the three basic discourse acts, or information packaging functions. As already highlighted by Croft et al. (2017) and Marneffe et al. (2021: 256–258), UD Relations are based to a certain extent on information packaging; in this specific case, modifying adjectives correspond to the 'amod' (adjectival modification)[[4]](#footnote-4) relation, while predicative adjectives are handled like verbs.

As for the semantic part of our comparative concept, we can exploit the UPOS layer by considering lexemes tagged as ADJ.

## Analytic case marker

In UD Relations, a distinction is made between lexical items acting as case markers i.e., adpositions, and lexical items involved in syntactic relations, i.e., arguments. For the purpose of the present study, we are interested in lexical items functioning as adpositions, which are addressed by the 'case' UD relation. Elaborating on Blake (2001): 1, the following definition is a general comparative concept for case markers, including adpositions, denoting every type of syntactic relation:

Case markers are linguistic devices used to signal that a noun stands in a syntactic relation with another linguistic element, serving as its head.

These linguistic devices can be syntactic, morphological or even of a phonological or prosodical nature. As we are dealing here with syntactic case markers, or ‘analytic case markers’ (Blake 2001: 9–12), we will not consider markers of morphological nature i.e., inflectional affixes, which are represented in Universal Dependencies as a Universal Feature[[5]](#footnote-5). Accordingly, we restrict our comparative concept to:

An analytic case marker is a lexical item used to signal that a nominal head stands in a syntactic relation with another linguistic element.

This comparative concept is still a bit too broad as we wish to exclude two syntactic relations happening at the noun phrase level, namely, possessive constructions and modification by means of other nouns.

Possessive constructions may be marked by analytic case markers such as English *of* and *’s*. However, their relations are labelled as 'nmod' or 'appos' UD Relations, and not using the 'case' UD relation. In other languages, possessive constructions are marked by inflection, which is part of UD Features, and not considered here (German genitive *s*, Dutch genitive *’s*).

The syntactic function of the analytic case marker corresponds to the 'case' UD Relation (Marneffe, Manning, et al. 2021: 269); one of the languages of our sample, Welsh, uses a specific relation subtype, 'case:pred', to address the particle *yn* that is found in nominal predicates[[6]](#footnote-6). For consistency with the other languages, we will not however consider this subtype. The corresponding lexical category of adpositions is captured by the ADP UPOS tag; since in all languages of our sample adpositions are closed category, we additionally combine the two UD annotation layers with lists of *lemmata* (see Appendix F).

## Article and Demonstrative

In Universal Dependencies, the 'det' relation is described as ‘the relation determiner (det) holds between a nominal head and its determiner’[[7]](#footnote-7) and is used (non-exclusively) whenever the modifier is marked with the DET UPOS tag. As mentioned in the main text, the DET UPOS tag varies wildly across languages, including word categories as different as articles, demonstratives, quantifiers, interrogative and personal/possessive pronouns. As a result, the 'det' relation and its UPOS counterpart, DET, not only appear particularly broad in their scope, conflating under a single category many types of syntactic relations, but are also cross-linguistically inconsistent.

Elaborating on Himmelmann (2001), we propose the following two comparative concepts for the article and the demonstrative:

Within a noun phrase, an article is a lexeme that occupies a fixed position and expresses certain features of the nominal head, namely: (in)definiteness and/or specificity; additionally, an article may also signal deictic and/or anaphoric reference of the nominal head it modifies.

Within a noun phrase, a demonstrative is a lexeme that may vary its position and functionally characterizes the nominal head for deictic and/or anaphoric reference.

The two comparative concepts differ on the position of the lexeme within the noun phrase; instances of the article have always a fixed position, while instances of the demonstrative may vary their position (Himmelmann 2001: 832). Among the languages of our sample, flexible demonstratives are found in Greek and Polish (see Sect. 4.2.5).

Furthermore, language-specific constructions of the demonstrative may involve the co-occurence of articles and of demonstratives (Himmelmann 2001: 840); one language of our sample, Welsh, shows such a strategy (King 2003: 28-34, 85-86).

The functional niche of the article is broader than of the demonstrative, in that “articles are distinguished from demonstratives by the fact that they can be used in semantic and pragmatic contexts in which demonstratives cannot be used”; the contexts in which an article may replace a demonstrative are however language-specific (Himmelmann 2001: 833).

Among the Universal features[[8]](#footnote-8), there is a parameter describing ‘(in)definiteness and/or specificity’, which is found on nouns, adjectives and articles. Unfortunately, some UD language models do not implement the feature, and deictic and anaphoric referenceis still missing from Universal Dependencies.

For these reasons, we cannot rely entirely on Universal Dependencies for the implementation of the two comparative concepts and, as with the analytic case marker, we employ language-specific lists of *lemmata*, along with the UD Relation and restrictions based on the UPOS layer.

Finally, a number of languages of our sample use 'det' subtypes; German, Italian, Polish and Portuguese use the 'det:poss' subtype[[9]](#footnote-9) for possessive determiners such as Italian *sua* ‘her/his’ in *casa sua* ‘her/his house’; English and Italian employ the 'det:predet' subtype[[10]](#footnote-10) to mark the determination of an already determined noun phrase (‘pre-determiners’, in that they are followed by a determiner), e.g., Italian *tutti* ‘all’ in *tutti gli anni* ‘all years, every year’. Since the above-described comparative concepts do not involve possessive determiners and pre-determiners, we are not concerned with these subtypes here.

# Appendix E: Frequency values of the word order patterns

The following table lists the frequency values of the three comparative concepts: analytic case marker (acm), relative clause, modifying adjective and of determiners; as in **Table 2**, frequency values are computed using the UD Relations annotation layer.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Language** | **acm** | | **rel. clause** | | **mod. adjective** | | determiners | |
|  | h-m | m-h | h-m | m-h | h-m | m-h | h-m | m-h |
| Danish | 4519 | 176541 | 38543 | 666 | 15326 | 79222 | 61 | 117720 |
| Dutch | 8915 | 182366 | 20667 | 50 | 3364 | 75492 | 78 | 189593 |
| English | 11224 | 194979 | 21691 | 7 | 1781 | 79274 | 9 | 175315 |
| French | 0 | 255628 | 33592 | 2 | 46208 | 38500 | 1 | 307612 |
| German | 392 | 164496 | n/a | n/a | 1290 | 81683 | 108 | 201215 |
| Greek | 33 | 141198 | 33216 | 7 | 3253 | 93618 | 5629 | 347286 |
| Italian | 38 | 238480 | 29216 | 0 | 48627 | 30817 | 0 | 272884 |
| Polish | 324 | 170968 | 21931 | 7 | 16025 | 68460 | 887 | 24177 |
| Portuguese | 1501 | 228272 | 28638 | 208 | 48101 | 22700 | 2295 | 273090 |
| Spanish | 470 | 264774 | 33412 | 0 | 52393 | 24555 | 207 | 288248 |
| Welsh | 81 | 35015 | 2749 | 7 | 8905 | 2315 | 339 | 16725 |

# Appendix F: Lists of lemmata

The following are lists of *lemmata* encoding in each language the analytic case marker, the article and the demonstrative comparative concepts. We have used the *lemma* annotation field whenever possible, querying for the uninflected (citation) form.[[11]](#footnote-11)

List of adpositions:

* Danish frequent colloquial/written adpositions: *af, angående, bag, blandt, efter, for, før, foran, forbi, foruden, fra, fraregnet, gennem, hos, i, iblandt, ifølge, igennem, imellem, imod, inden, indtil, langs, med, mellem, mod, nær, om, omkring, over, på, per, siden, til, uden, under, undtagen, ved, vedrørende, via* (Lundskær-Nielsen and Holmes 2010: 417-430, 486-487). Additionally, some complex adpositions: *fremfor, henad, henimod, ovenpå, udenfor*.
* Dutch frequent colloquial/written adpositions: *à, aan, aangaande, achter, achteraan, achterin, achterom, achterop, achteruit, ad, af, alvorens, behoudens, beneden, benevens, betreffende, bezijden, bij, binnen, blijkens, boven, buiten, circa, conform, contra, cum, dankzij, de, door, ex, exclusief, gedurende, gehoord, genstaande, getuige, gezien, halverwege, hangende, in, inclusief, ingevolge, inzake, jegens, krachtens, langs, met, middels, na, naar, naast, namens, om, omstreeks, omtrent, ondanks, onder, ongeacht, onverminderd, op, over, overeenkomstig, per, pro, qua, rond, sedert, sinds, staande, te, tegen, tegenover, teneinde, tijdens, tot, trots, tussen, uit, van, vanaf, vanuit, vanwege, versus, via, volgens, voor, vooraan, voorbij, voorin, voorom, voorop, vooruit, wegens, zonder* (Broekhuis 2013: 30-34, 40).
* Dutch also has a long list of complex adpositions (Broekhuis 2013: 29), including: *binnendoor, binnenin, bovenaan, bovenop, doorheen, middenin, naartoe, omheen, onderaan, onderdoor, onderin, onderlangs, onderop, onderuit, overheen, tegenaan, tegenaan, tegenop, tegenover, tussendoor, tussenin, tussenuit, vanachter, vandaan, vanonder.*
* Additionally we include *heen*, which can be part of a circumposition; *Jan sprong over het hek heen* (Broekhuis 2013: 49). We also include *terug*, which is usually a particle but which can be a postposition; *Jaren terug ben ik daar ook geweest* (Broekhuis 2013: 34).
* English: *aboard, (a)bout, above, absent, according, across, adjacent, after, against, ago, along, alongside, amid(st), among(st), apart, apropos, around, aside, at, away, back, bar, (be)cause, before, behind, below, beneath, beside, between, beyond, by, close, consequent, contrary, despite, down, due, during, effective, ere, close, except, exclusive, failing, far, following, for, from, given, in, including, inside, instead, into, irrespective, like, near, nearby, next, notwithstanding, of, on, onto, opposite, out, outside, over, past, per, preliminary, preparatory, previous, prior, pursuant, regarding, regardless, round, ’s, since, subsequent, through, throughout, till, to, together, toward(s), under, underneath, unlike, until, unto, upon, via, wanting, with, within, without* (Huddleston and Pullum 2002: Chapter 7).
* We do not include here all the adpositions that are homonymous with the gerund-participle (Huddleston and Pullum 2002: 611), as most of them do not appear in our data (this may point to an issue with UD parsing). The same goes for most of the adpositions formed on the basis of *here-*, *there,* and *where*-. We also leave out many adpositions that do not take NP complements.
* French*: à, avec, après, avant, chez, comme, concernant, contre, dans, de, debout, dehors, depuis, derrière, devant, durant, en, entre, envers, hors, moyennant, par, parmi, pendant, pour, près, sans, sauf, selon, suivant, sur, vers* (Batchelor and Chebli-Saadi 2011: 113-117, 564-566, 569, 571-605).
* German: *ab, abseits, an, an, angesichts, anhand, anstelle, auf, aufgrund, aus, außer, außerhalb, bei, beiderseits, bezüglich, binnen, bis, dank, darunter, diesseits, durch, einschließlich, empor, entgegen, entlang, entsprechend, fernab, für, gegen, gegenüber, gelegen, gemäß, halber, hin, hindurch, hinsichtlich, hinter, in, infolge, inklusive, inmitten, innerhalb, jenseits, kraft, längs, laut, links, mangels, minus, mit, mitsamt, mittels, nach, neben, obengenannter, oberhalb, ohne, per, plus, pro, rechts, ringsum, seit, seitens, seitwärts, statt, trotz, über, um, unter, unterhalb, unweit, von, vor, während, wegen, zeit, zu, zufolge, zugunsten, zusammen, zwecks, zwischen* (Eisenberg and Schöneich 2020: 200, Breindl 2020d).
* Greek: από, για, εις, εκ, εκτός, εν, ένεκεν, εντός, επί, κατά, κατόπιν, μαζί, με, μετά, μέχρι, μέσω, μεταξύ, πριν, προ, προς, χάριν, σε, ως (Holton, Mackridge, and Philippaki-Warburton 2012: 460-499).
* Italian: *a, accanto, addosso, assieme, attorno, attraverso, circa, come, con, contro, da, davanti, dentro, dietro, dinnanzi, dirimpetto, dopo, durante, eccetto, entro, esterno, fino, fra, fuori, incontro, indietro, innanzi, insieme, intorno, lontano, lungo, neanche, nemmeno, nonostante, oltre, per, perfino, persino, presso, prima, riguardo, rispetto, salvo, secondo, senza, sino, sopra, sotto, su, tra, tramite, tranne, traverso, verso, vicino* (Maiden and Robustelli 2013: 174-179, 185-187).
* Polish: *bez, blisko, dla, do, dokoła, dookoła, dzięki, koło, ku, między, mimo, na, nad, naokoło, naprzeciw, naprzeciwko, o, obok, od, ode, oprócz, po, pod, podczas, podług, pomiędzy, pomiendzy, poniżej, pośród, pośrodku, powyżej, poza, prócz, przeciw, przeciwko, przed, przez, przezyć, przy, spod, spode, spomiędzy, spośród, spoza, sprzed, u, w, wbrew, według, wewnątrz, wobec, wokół, wokoło, wskutek, wśród, z, za, znad, zza* (Swan 2002: 341-373).[[12]](#footnote-12)
* Portuguese: *a, ante(s), até, de, depois*. The following adpositions are reported by the grammar but are not lemmatized by the parser, or occur with ackward forms and/or very few occurences: *após, com, conforme, contra, em, excepto, desde, durante, entre, fora, mediante, para, por perante, salvo, segundo, sem, sob, sobre, visto* (Whitlam 2011: 176-186).
* Spanish: *a, ante(s), aparte, bajo, cabe, como, con, contra, de, delante, depois, desde, detrás, durante, en, encima, entre, excepto, gracia(s), hacia, hasta, in, incluso, mediante, para, por, respecto, salvo, según, sin, sobre, tras, via* (J. Butt, Benjamin, and Rodríguez 2019: 480).
* Welsh: *â, ag, am, ar, argyfer, arngylch, at, cyn, dan, dros (tros, thros, drost), efo (hefo), er, erbyn, fel, gan, ger, gyda, gyferbyn, (am) gylch, heb, heblaw, heibio, hyd, i, megis, mewn, (er) mwyn, nes, o, oddi (ar), oddiwrth (oddi wrth), ogylch, oherwydd, rhag, rhwng, tan, trwy (drwy), tua, wedi, wrth, ymhlith, yn, ynghylch* (King 2003: 273-301, Williams 1980: 127-130, 143).

List of articles: ? = partitive / indefinitive plurals; (forms) = inflected forms

* Danish: *en, et* (Lundskær-Nielsen and Holmes 2010).
* Dutch: *de, het, een* (own knowledge).
* English: *the, a* (own knowledge).
* French: *le, un* (*l’, la, les, d’? , de?, une)* (Batchelor and Chebli-Saadi 2011: 101-112).
* German: *ein* (*eine, einem, einen, einer, eines*), *der* (*dem, den, des*) *die* (*der*), *das* (*dem*, *des*), *die* [PL] (*den, der*) (Breindl 2020a, Breindl 2020c).
* Greek: ο, ένας (η, οι, τα, τη, την, της, τις, το, τον, του, τους, τω, των, ένα, έναν, ένας, ενός, μια, μιας) (Holton, Mackridge, and Philippaki-Warburton 2012: 52-53).
* Italian: *il, uno* (*gli, le, lo, l’, una, un’, un, dei?*) (Maiden and Robustelli 2013: 61, 79-80).
* Polish. There are no articles in Polish.
* Portuguese: *o, um* (*os, a, as, uns?, uma, umas?*) (Whitlam 2011: 32, 35-37).
* Spanish: *el, uno* (*la, las, los, un, una, unas?*, *unos*?) (J. Butt, Benjamin, and Rodríguez 2019: 30, 45, 50-52).
* Welsh: *y* (*yr, ’r*). There is no indefinite article in Welsh (King 2003: 29).

List of demonstratives:

* Danish: *denne* (*her*), *dette* (*her*), *disse* (*her*), *den her, det her, de her, den* (*der*), *det* (*der*), *de* (*der*) (Lundskær-Nielsen and Holmes 2010: 208).
* Dutch: *die, dit, dat, deze* (own knowledge).
* English: *this, that (those, these)* (own knowledge).
* French: *ce* (*ces, cet, cette*) (Batchelor and Chebli-Saadi 2011: 610-612).
* German: *dies* (*diese, diesem, dieser, dieses*) (Breindl 2020b).
* Greek: αυτός, εκείνος, τούτος (αυτά, αυτές, αυτή, αυτήν, αυτής, αυτό, αυτοί, αυτόν, αυτός, αυτού, αυτούς, αυτών, εκείνα, εκείνες, εκείνη, εκείνο, εκείνοι, εκείνον, εκείνος, εκείνου, εκείνους, εκείνων, τούτη, τούτο, τούτον, τούτους) (Holton, Mackridge, and Philippaki-Warburton 2012: 400-403).
* Italian: *codesto, quello, questo* (*codeste, codesti, quegli, quei, quel, quell’, quella, quelle, quello, quest’, questa, queste, questi*) (Maiden and Robustelli 2013: 81-83).
* Polish: *ten* (*to tego, temu, to, tym, ta, ci, te, tej, tych, tà, tymi*), *ów* (*owego, owemu, owym, owo, owa, owej, owi, owe, owych, owymi*) (Swan 2002: 171-174).
* Portuguese: *este, esse, aquele* (*esta, essa, aquela, estes, esses, aqueles, estas, essas, aquelas*) (Whitlam 2011: 70-72).
* Spanish: *este, ese, aquel* (*esta, estas, esto, estos, esa, esas, eso, esos, aquella, aquellas, aquello, aquellos*) (J. Butt, Benjamin, and Rodríguez 2019: 85).
* Welsh: *hwn, hwnnw, hon, honno, hyn, hynny* (King 2003: 85).

1. These categories can be further analysed in terms of semantic properties, as shown by Croft (2001: 87). [↑](#footnote-ref-1)
2. <https://universaldependencies.org/u/dep/acl.html> [↑](#footnote-ref-2)
3. <https://universaldependencies.org/u/dep/acl-relcl.html> [↑](#footnote-ref-3)
4. <https://universaldependencies.org/u/dep/amod.html> [↑](#footnote-ref-4)
5. <https://universaldependencies.org/u/feat/Case.html> [↑](#footnote-ref-5)
6. <https://universaldependencies.org/cy/dep/case-pred.html> [↑](#footnote-ref-6)
7. <https://universaldependencies.org/u/dep/det.html> [↑](#footnote-ref-7)
8. <https://universaldependencies.org/u/feat/Definite.html> [↑](#footnote-ref-8)
9. <https://universaldependencies.org/u/dep/det-poss.html> [↑](#footnote-ref-9)
10. <https://universaldependencies.org/u/dep/det-predet.html> [↑](#footnote-ref-10)
11. Greek, Portuguese and Welsh articles and demonstratives are not lemmatized in the UD models we have used for parsing, so we have to query for individual word. [↑](#footnote-ref-11)
12. Thanks to Ekaterina Lapshinova-Koltunkski and Krzys Koltunski for checking this list for us. [↑](#footnote-ref-12)