Quid verbumst?
Applying a definition of word to Latin in Universal Dependencies

Flavio Massimiliano Cecchini
KU Leuven – Erasmushuis, Blijde-Inkomststraat 21, 3000 Leuven, Belgium
flaviomassimiliano.cecchini@kuleuven.be

Abstract

Words, more specifically "syntactic words", are at the centre of a dependency-based approach like Universal Dependencies. Nonetheless, its guidelines do not make explicit how such a word should be defined and identified, and so it happens that different treebanks use different standards to this end. To counter this vagueness, the community has been recently discussing a definition put forward in (Haspelmath, 2023) which is not fully uncontroversial. This contribution is a preliminary case study that tries its hand at concretely applying this definition (except for compounds) to Latin in order to gain more insights about its operability and groundedness. This is helped by the spread of Latin over many treebanks, the presence of good linguistic resources to analyse it, and a linguistic type which is probably not fully considered in (Haspelmath, 2023). On the side, this work shows once more the difficulties of turning theoretical definitions into working directives in the realm of linguistic annotation.

1 Introduction

The notion of (syntactic) word is at the centre of a dependency-grammar approach like that of Universal Dependencies (UD; de Marneffe et al., 2021):[1] quoting from the project's guidelines, "dependency relations hold between *words* [...] the basic units of annotation are *syntactic* words". How this fundamental unit of morphosyntactic analysis can or needs to be defined, however, is still at the centre of debates and contrasting opinions in the community, and it is not directly confronted by the guidelines. In this sense, as part of the COST action "UniDive: Universality, Diversity and Idiosyncrasy in Language Technology" (Savary et al., 2024),[2] task 2.1 of the working group 2 (lexicon-corpus interface) has been devoting itself to the harmonisation of the definition of syntactic word across languages, gathering information through surveys and regular meetings, and presenting some results at the 3rd general meeting in Budapest in 2025. The main starting point for this endeavour has been established in a paper by Haspelmath (2023), where a clear-cut definition of word is given, based among others on his previous works on this topic (Haspelmath, 2017, 2021). The same author admits that his definition might result "unnatural" (Haspelmath, 2023, cf. §4.5), as it does not precisely overlap with how words have been traditionally identified by each respective language. In fact, even inside UD and UniDive's communities, but not limited to them, some approaches are being brought forth to simply "go beyond" the notion of word, cf. (Haspelmath, 2025) or the 1st proposed shared task on morphosyntactic parsing.

The definition of word is a particularly acute issue for Latin in UD, as this language counts many treebanks managed by different teams and originating from different annotation standards. This fact, among other things, affects the notion of wordhood in each of these treebanks, so much so that efforts have already taken place in order to deal with Latin's morphosyntactic harmonisation (Gamba and Zeman, 2023b,a). The present paper wants to offer a more in-depth overview about "what is a word"[3] for what concerns Latin, tapping from the experience of UniDive working group 2 task 2.1's survey, and moving along two main lines: on the one hand, sketching, on the basis of UD-treebank-driven empirical observations, what the characteristics of the "Latin word" are, and, on the other hand, putting to the test Haspelmath's (2023) framework. We believe that

[1] https://universaldependencies.org
[2] https://unidive.lisn.upsaclay.fr
[3] Hence the title *quid verbumst* 'what is a word', with the univerbation *verbumst* for *verbum est*, cf. Section 7.

such a case study can provide useful insights to its actual operability, both for the harmonisation of Latin treebanks and of all treebanks in general. Latin has the advantage of presenting a good amount of data in UD, of possessing many valid resources for its linguistic analysis, and also of bearing some typological characteristics which we feel are not fully taken into consideration in (Haspelmath, 2023). In this study we only refrain from tackling compounding, as we deem the scope of that phenomenon to be worthy of another entire, dedicated study, which can profit from the definition of "simple" words first.

Further, despite being centered on Latin, this work makes scripts available (see Section 3) which can be directly, or with little adjustments, applied to any other UD treebank, in order to replicate most of the results shown here for other languages. In fact, it is our hope that the scope of this paper[4] be widened and joined by similar case studies so as to create an as cross-linguistic as possible framing of Haspelmath's (2023) definition. UD is a collective endeavour and we contribute to it with the language we know best in its context: Latin.

Section 2 constitutes a brief linguistic profile of Latin, while Section 3 briefly describes the current situation of Latin in UD and the data used in this study; Section 4 comments on Haspelmath's (2023) definition of word; Sections 5 and 6 bring forth the analysis; Section 7 makes some final remarks and Section 8 concludes.

2 Latin: the language

We supply here a concise linguistic profile of Latin in order to make the following discussion easier to follow for those not well acquainted with this language. From the historical point of view, the Latin language emerged from the Latino-Faliscan branch of Italic, as spoken by the tribe of *Latini* (Latins) in *Latium*, roughly corresponding to the modern-day *Lazio* region of Central Italy.[5] As such, it is a typical, as it were, ancient Indo-European language showing an extensive fusional morphology in all major word classes (nominals, modifiers, verbs).[6] Inflection of nominals (including "nouny" adjectives/determiners, cf. Stassen, 2003, §9.1, §15) takes place through suffixes encoding at the same time case, grammatical gender and number, and is traditionally subdivided into five "declensions" for nouns and two "classes" for adjectives according to the similarities between sets of suffixes (concerning a "thematic vowel" or the absence thereof); as discussed in Sections 5 and 6, it is quite rare to find a nominal element without any inflectional affixes. Verbs (including auxiliaries) also inflect mostly fusionally for aspect, mood, number and person of the subject, tense, and voice, and can also take on nominal forms with their own paradigms (cf. Cecchini, 2021 for an overview from the point of view of UD). Verbal inflection is traditionally subdivided into four main "conjugations" (again according to the presence or absence of a "thematic vowel"), but aspectual distinctions are often much more complex from a morphological point of view (see e. g. Pellegrini, 2023), and this might be a consequence of the whole verbal system shifting from being aspect-based to tensed in preliterate times (cf. again Stassen, 2003, §10.2). Word derivation is, just like inflection, predominantly suffixal, especially with regard to changes in word classes (e. g. deverbal nouns etc.), while prefixation seems to be restricted to the expression of more "lexical" categories such as *Aktionsart* (cf. Haverling, 2000 for the verbal system) and degree. This all correlates well with the original dominant verb-final, more precisely SOV, word order of Latin, which however seems to already show the signs of a shift towards the later Romance SVO order in historical times (for example, adpositions are chiefly *pre*positions; cf. Adams, 1976), and can vary noticeably for pragmatic and information-packaging reasons anyway. For finer-grained details on Latin syntax, and much more, we refer to the state-of-the-art, comprehensive, monumental work by Pinkster (2015, 2021).

Latin's presence following UD's annotation formalism is currently[7] articulated over six treebanks in UD proper, and at least two other relevant resources, as shown in Table 1 (where the figures for LASLA do not take into account the works already

[4] With some reason defined by a reviewer a "position paper".

[5] Comprehensive references about this topic in general and for the relation of Latin with other Italic languages are (Clackson and Horrocks, 2007), and the Introduction of (de Vaan, 2008).

[6] Under the class of modifiers, what are traditionally called adverbs are problematic in that they are systematically treated as distinct uninflectable lexemes even when they are transparently and productively derived from some base (Cecchini, 2024).

[7] As of v2.15, released in November 2024 (Zeman et al., 2024).

Name/Code	References	Syntactic words	Syntax	InflClass
CIRCSE	(Iurescia et al., 2024)	18 968	Yes	Yes
IT-TB	(Cecchini et al., 2018; Passarotti, 2019)	450 517	Yes	Yes
LLCT	(Cecchini et al., 2020a)	242 411	Yes	Partially
Perseus	(Bamman and Crane, 2011)	29 221	Yes	No
PROIEL[8]	(Haug and Jøhndal, 2008; Eckhoff et al., 2018)	205 566	Yes	No
UDante	(Cecchini et al., 2020b)	55 519	Yes	Yes
Sabellicus	(Gamba and Cecchini, 2024)	10 755	Yes	Yes
LASLA[9]	(LAS, 2024)	1 820 405	No	Yes

Table 1: Overview of the annotated resources for Latin in UD or following UD's formalism used in this study. Syntactic words include punctuation marks.

3 Latin: the data

included in CIRCSE).

Taken together, they reach 1 002 202 syntactic words only in UD (making Latin one of the most represented languages therein, coming out 9th out of more than 150) and 2 833 362 in total. The language in these annotated corpora, despite its extended chronology (from Antiquity to the Renaissance), is rather homogenous, if not for style and genre, from the morphosyntactic point of view, given how it is mostly oriented towards the Classical variety (cf. Pinkster, 2015, §1.7 for this periodisation) of the 1st c. BCE (Clackson and Horrocks, 2007, chs. V-VI), or at least it does not diverge too much from it (with the exception perhaps of LLCT, cf. Korkiakangas and Passarotti, 2011, §3). Their annotation, though, is not always homogeneous, as in some cases it is natively UD, in others the result of conversions from other annotation standards, but has been harmonised over time (see Gamba and Zeman, 2023b,a); in the case of LASLA, the original annotation does not even include syntax. With respect to our case study (Sections 5 and 6), we are also interested in the annotation of inflectional classes (UD's morpholexical feature `InfClass`), which is not always present.

Some technical remarks for the analyses in later Sections: whenever an analysis is performed, it is always intended to take place on the subcorpus formed by those resources on which it can fully apply (e. g. excluding LASLA if syntax is concerned); all lemmas and forms are normalised (in particular: lowercased, $v > u$, $j > i$); to avoid data sparsity, parts of speech have been reorganised partly following (Cecchini, 2024), in particular `PROPN` is subsumed under `NOUN`,[10] `NUM` under `DET`,[11] `ADV` redistributed according to the parts of speech of their bases;[12]; `DET`, `NUM` and `PRON` are considered synsemantic (or "functional", or "grammatical") word classes; by lexeme we mean all forms described by a unique couple of lemma plus part of speech.[13] All data cited in this work and the Python scripts to produce them (for Latin, but adaptable to any other UD language) are made available through a sharing platform,[14] mentioned in what follows as "the repository". Some are however shown in Appendix A for ease of reference.

4 Defining the word: MH and UD

Referring to (Haspelmath, 2023) for the complete discussion, we only highlight some main points here. Haspelmath's (2023), henceforth MH, and UD's frameworks use transversal, but partly overlapping and correlated categories. MH's definition is based on three morphological objects which represent themselves word types, i. e. free morphs, clitics, and roots, to which compounds are added as "second-order" formations. These objects are themselves based on the more fundamental notion of morph (Haspelmath, 2020), and on four combining fundamental properties, which we resume in our

[8] `http://dev.syntacticus.org/proiel.html`

[9] `https://www.lasla.uliege.be/cms/c_11821932/fr/lasla-lasla-dataverse`

[10] Cf. guidelines: "A proper noun is a noun (or nominal content word) that is the name (or part of the name) of a specific individual, place, or object. [...] Note that `PROPN` is only used for the subclass of nouns".

[11] Cf. guidelines: "Other words functioning as determiners".

[12] See list at `https://github.com/Stormur/OrderlyAdverbs/blob/main/Latin/ADV_omnia.tsv`. Note that we group `REL` with `PRON`, following the most conservative proposal in (Cecchini, 2024, §5.1.5).

[13] This still does not eliminate some further ambiguities, but it is a good enough approximation to identify morphosyntactic patterns.

[14] `https://github.com/Stormur/quidverbumst`

	Contentful	Bound	Selective	Affixes
Free morph	-	No	No	No
Clitic	-	Yes	No	No
Root	Yes	-	No	-

Table 2: Morphological objects in the definition of word by Haspelmath, 2023, and their properties, if specified.

words as: contentfulness, boundness, selectivity[15] and presence of affixes. The relation between all these elements is summarised in Table 2. The notion of part of speech, a tenet of UD (de Marneffe et al., 2021, §2.2.2), is not explicitly mentioned by MH, but it is hinted at by his description of a root as "a morph denoting an action, an object or a property", which mirrors the subdivision of words into phrasal types at the core of UD (de Marneffe et al., 2021, §2.1.1); moreover, the notion of contentfulness is parallel to UD's distinction along the grammaticalisation cline between autosemantyc and synsemantic words (de Marneffe et al., 2021, §2.2.1), codified into pairs of part-of-speech tags sharing the same phrasal type, such as the modifiers `ADJ` vs. `DET`. In contrast, the notions of boundness and selectivity are not explicitly considered in UD's formalism. Since we are rooting our study in UD's formalism, we are abiding by its classification of words into parts of speech, and we are interested in seeing how these can be mapped onto MH's word types. In fact, we have to remark that UD's analysis does not reach the level of individual morphs,[16] so that we have to approximate MH's framework at the level of lexemes, justified by the circumstance that the grouping into lexemes is (usually) centered around commonality of morphs. In MH's scheme, roots are actually generalised by and inferred from what he calls free forms, which possibly include other non contentful (= synsemantic) morphs, not necessarily just affixes. At first, free forms seem to overlap with the notion of sentence in UD, and so with that of main clause with no subordinates. In the following, we will try investigate what these objects look like and to what extent they are applicable for Latin.

5 *Formae liberae*, or Latin free forms

The fact that MH also considers elliptical free forms poses a first serious difficulty for us with respect to their identification in our data: if we assume that any argument of a clause can be "extracted" from it to form an elliptical version thereof, we have to consider all arguments in all sentences, and recursively any subordinated arguments of such arguments, not just whole sentences, as possible free forms. This is problematic to deal with both computationally and definitionally. For this reason, we decide to exclude elliptical constructions and limit our first analysis to free forms appearing as sentences coinciding with explicit simple clauses. Technically, we consider any sentence for which the subtree of its `root`[17] does not contain any edge labelled with a clausal relation[18] (in other words, we exclude sentences with subordinate clauses of any kind), and where all nodes except the `root` belong to synsemantic parts of speech,[19] thus excluding interjections (`INTJ`) as extra-lexical elements, unannotated words (`X`), and ignoring non-lexical elements (`PUNCT`, `SYM`). We ignore `parataxis`, and exclude sentences presenting a co-ordination (relation `conj`) at the `root`, to avoid elliptical structures: ellipsis is a pervasive phenomenon in similar constructions, with no clear preference in Latin for being left- or right-headed,[20] and it is unfortunately not (yet) signalled in UD's basic annotation. For the same reason, we exclude sentences presenting an `orphan` relation, which is the hallmark of ellipsis. We also require the feature `VerbForm` with value `Fin` to appear in the clause,[21] or, to accommodate possible "zero copulas", that there be non-clausal arguments (apart from `orphan`) or predicate modifiers.[22] Finally, words related by the "headless" relations `fixed` and `flat` are

[15] "cannot occur on roots of different root classes."

[16] Despite efforts to define and operationalise an annotation in this sense, see e. g. (Gamba et al., 2024).

[17] We use the notation for dependency relations to distinguish UD's sense of root from MH's.

[18] These are `csubj`, `ccomp`, `xcomp`, `advcl`, `acl`.

[19] These are `ADP`, `AUX`, `CCONJ`, `DET`, `PART`, `PRON`, and `SCONJ`.

[20] Compare, in the same text in UDante, a left-headed ellipsis in `DVE-89` and a right-headed one in `DVE-313`.

[21] This requirement is partly Latin-specific, as main clauses are always headed by a finite predicate, or at least annotated this way following grammatical tradition.

[22] These are `nsubj`, `obj`, `iobj`, `obl`, `vocative`, `dislocated`, `advmod`, `discourse`.

considered together as single units.[23] Contrary to MH's definition, and expanding it, we do not require the root to be autosemantic, for reasons discussed later.[24]

Out of a total of 60 194 syntactically annotated sentences in our corpus, 50 355 do not present a root co-ordination, and among these we find 1410, the 2.3% of all sentences, distinct (normalised) such free forms, headed by 606 distinct lexemes (for a lexeme/form ratio of 43.0%). Figure 1 (above) in Appendix A shows the distribution of the latter with respect to their parts of speech. As expected, we see that UD's main autosemantic parts of speech VERB (1001 types, 350 lexemes), NOUN (169, 118) and ADJ (142, 94) dominate: in fact, their definitions greatly overlap with those of MH's roots, as noted in (Haspelmath, 2023, p. 287), even if UD does not make semantic distinctions about the concreteness of their meanings, so that these classes will include more morphs in UD than in MH's scheme. We also cannot draw immediate conclusions about the boundness of these roots themselves: they are surely non-selective, in that they are independent from the actual realisations of their arguments and also from their linear order (at least in Latin this being only determined by pragmatic factors, cf. Spevak, 2010), but some of them, surely not all (cf. the single forms discussed below), might actually require some arguments to be expressed, e. g. a direct object. Such an investigation is beyond the scope of this paper, and here we can just note down that boundness emerges as an orthogonal property to UD's parts of speech VERB, NOUN, ADJ and ADV. There is a quantitative gap between these and synsemantic parts of speech, which can be partly explained with the more closed nature of the latter. The exiguity of ADV, besides their strong relational and metapredicative nature,[25] and of AUX, not expected to head a predicate by definition,[26] can also explain their positions after PRON and DET, this order otherwise being symmetrical to that of VERB-NOUN-ADJ.

An even greater divide beckons in the synsemantic field between PRON, DET and AUX, and the other parts of speech not appearing in even one of our free forms:[27] ADP, CCONJ, PART and SCONJ. This is also to be expected, because, beyond being synsemantic, these parts of speech are more or less explicitly defined as connectors, as exclusively grammatical elements introducing nominal or clausal arguments, or conveying morphological or pragmatic features (PART), so that it is not easy to imagine them as predicate heads. Under this light, we can consider the elements of these classes to be bound to the heads of the phrases they accompany, and so, passing to MH's framework, they are the prime candidates for clitics. In the Latin system, these words are also set apart from all others by a lack of inflectional affixes, that is, they are unique forms in their paradigms, even if this does not necessarily mean that they are bare morphs (see Section 6). But there is an even more striking characteristic that strengthens their identification with clitics: phonologically, many such words present endings which are not shared, or only marginally, by autosemantic classes, remarking their prosodic boundness. It is the case of ADP *sub* 'under', CCONJ *sed* 'but', PART *non* 'not' (Cser, 2020, §2). The interesting consequence is that Latin seems to have morphologically codified the distinction between roots and clitics, with intermediate degrees for non-clitic synsemantic elements (see Section 5.1): the question then arises how common this is typologically. In any case, the identification of roots (if we loosen the request for a concrete meaning) with VERB, NOUN, ADJ and possibly, partly ADV (intended *stricto sensu* "true", contentful adverbs such as *saepe* 'often' or *uix* 'with difficulty', and possibly relators like *supra* 'above') on one side, and of clitics with ADP, CCONJ, PART and SCONJ on the other side plausibly looks to be universal.

[23] Though the impact of this is extremely limited in our data.

[24] The results are presented in the repository as the file `freeforms.tsv`.

[25] Cf. (Cecchini, 2024, §4.2). Further, we notice that the two extracted free forms (all in ITTB, `dev-s530`, `dev-s533` and `train-s9004`) are actually faultily segmented subordinates of a previous sentence, and present ellipsis: so e. g. *quia non semper est* '[. . .] since there is not always [any]', where the existential AUX *est* should be promoted to head.

[26] We notice indeed that out of the 18 extracted free forms headed by the AUX *sum* 'to be', most come from PROIEL and should be annotated otherwise, cf. (Gamba and Zeman, 2023b, p. 11), and there are issues of sentence segmentation and ellipsis. The only viable free form would be *esto* '(so) shall it be' (PROIEL 78649).

[27] The single occurrence of ADP *usque septies* 'up to seven times' (PROIEL 13811) should actually have *septies*, a multiplicative numeral, as its head instead of *usque* in the current annotation standard. The two occurrences of PART are also due to faulty annotations.

5.1 Not roots nor clitics

The position of the roots' synsemantic counterparts AUX, PRON and DET, in MH's system is less clear. Only to AUX, given its extremely limited appearance in clausal free forms, it comes easy to associate the status of clitic on the road to an affix, at least for Latin (and observed univerbation practices strengthen this claim, cf. Lehmann, 2020 about this process), when looking at its boundness, even though the verbal inflection of this class is a problem if clitics have to be bare morphs. The same argument could actually be repeated for ADV, in the sense that its members can only be associated to a predicate, and not head a predicate themselves, making them bound, but again, possibly bearing (derivational) affixes. It has to be admitted, though, that this conclusion is subject to the current annotation practices for ADV, which are far from being uniform or coherent (cf. Cecchini, 2024 and Section 3). As for PRON and DET, strictly speaking, they should not be counted as words in MH's system, as they are no roots, and in general not even free morphs (see Section 6). They are not affixes either, as their distribution is parallel to that of NOUN and ADJ, in the sense that they are no more selective than those. This grey zone seems to arise from what appears to be a mixing of purely distributional, i. e syntactic (boundness and selectivity) and semantic (contentfulness) criteria in MH's definition of word types, in particular the request of a "concrete" meaning for roots, criteria which however look to be orthogonal to each other. In other words: AUX, PRON and DET do not seem to distribute that much differently than their autosemantic counterparts, but, at least in Latin, they cannot be considered words according to MH's framework simply because their meanings are more abstract and less complex, but they happen to be inflected; at the same time, a good candidate for root, the contentful subset of ADV, has a more clitic-like behaviour than them. Section 7 will address this point further, after having a look at the distribution of affixes in Latin.

6 *Liberi morphoi*, or Latin free morphs

In MH's framework, a particular subclass of free forms are free morphs, i. e. free forms consisting of only one morph. Among the clausal free froms discussed in Section 5, only members of VERB (and in one case also AUX) appear as standalone forms (119 types, 97 lexemes), e. g. *obsecra* 'implore' (PROIEL 50017), but these forms usually do not consist of single morphs. Given our chosen constraints, we do not observe other standalone representants of non-clitic parts of speech there. However, even if an explicit copula, so at least one extra morph, appears to be the most frequent strategy to from predicates for non-verbs (we record this 78.7% of the times for NOUN, 79.6% for ADJ, 62.3% for PRON and 78.9% for DET), zero copula is indeed possible in Latin, cf. (Stassen, 2003, p. 676),[28] as are other kinds of one-word utterances. Thus, in search for free morphs we can shift our attention to sentences consisting of only one syntactic word (barring the presence of alexical PUNCT and SYM), allowing us to broaden the analysis to our whole Latin corpus; the distribution of such single-word free forms across parts of speech is then also shown in Figure 1 (below) in Appendix A.[29] The picture is very similar to that of clausal free forms: beyond noise in the data, we see that DET, PART and AUX are relatively more represented, but, more interestingly, INTJ is now more relevant, with occurrences of e. g. *st*, *hem* and *attatae*. This is surely due to the contribution of LASLA, featuring comedies and more everyday language (also shown by the appearance of vocatives among nouns, 10 out of 115 form types), but it does highlight a class of utterances which cannot be interpreted as predicates, and whose members often have an unclear lexical status. They are in fact unanalysable forms, and we can take them as free morphs; incidentally, an English interjection (*ouch*) is also among MH's examples for this class. However, the key verdict is that, for Latin, free morphs substantially begin and end with interjections.

It is often stated that Latin has a rich morphology, and this can actually be quantified with the help of the available lexical resources. Using LatInfLexi (Pellegrini and Passarotti, 2018; Pellegrini, 2024), we can assess, for what regards VERB and NOUN, that only a small subclass of NOUN lexemes admits bare morphs as parts of their paradigms, e. g. *mel* 'honey', *sōl* 'sun', or *uir* 'man' (all having concrete meanings).[30] These forms can be considered the stems[31] of their respective paradigms, since

[28] Even if it is very hard to pinpoint and distinguish from elliptical structures in the current UD's annotation standards.

[29] Complete results are in the file `singleforms.tsv` in the repository.

[30] The complete list is in the file `radicalforms.tsv` in the repository.

[31] Admitting here for semplicity and without loss of gener-

all others are obtained by suffixation (which is predominant in Latin), but this does not make them unmarked for grammatical categories: in fact, Case (nominative, and also accusative for *mel*), Number (singular), and partly Gender are completely predictable from their forms.[32] We identify 81 out of 1038 noun lexemes of this kind in LatInfLexi; however, free morphs among these are effectively only a quarter, as many are derived, e. g. *ēruptio* 'a breaking out', an action name form from VERB *ērumpo* 'to break forth', itself bearing the preverb *ē* 'out of'.[33] Thus, the magnitude of this phenomenon appears almost irrelevant for Latin, and even more so as we do not find any of similar free forms in our corpus. Two related, parallel questions arise here: whether there exist non-clitic lexemes bearing no inflectional, at most derivational (in MH's terms, "not required") affixes, i. e. uninflectable words; and, even if inflectable, how many underived lexemes, i. e. admitting at most inflectional (in MH's terms, "required") affixes, make up the Latin lexicon.

For the first question, we rely on the annotation of the feature InflClass 'inflectional class' in some of the Latin treebanks. From these, we extract all the lexemes which possess forms that are never annotated for an inflectional class.[34] Of the 2175 identified lexemes, we focus on VERB, NOUN, ADJ, their synsemantic counterparts AUX, PRON, DET, and ADV. After having excluded abbreviations and symbolic numerals, and filtered out some dubious or simply erroneous annotations through manual inspection,[35] we gather only a handful of contemplatable cases, that is, by their parts of speech:

NOUN *māne* 'morning', *here* 'yesterday', *crās* 'tomorrow'

ADJ *nēquam* 'worthless', *satis* 'enough', *uolup* 'pleasant'

PRON *quandō* 'when', *ubī* and *ibī* 'where', *unde* 'whence', *cūr* 'why', *nīl* 'nothing'

DET *ita* 'so', *tam* 'as much', *tot* and *quot* 'as many', *siremps* 'same', and derived forms; cardinal numerals above 3

ADV many "true adverbs" such as *ferē* 'approximately'; relators such as *suprā* 'above'

It is interesting to notice some patterns, especially about quantities and time indications, and that modifiers (ADJ, ADV, DET) appear to be slightly more represented than the other types, but the numbers are vanishingly small, and this group by all means constitutes an "inflectionally closed class".

To answer the second question, we turn to Word Formation Latin (WFL, Litta and Passarotti, 2019). From all the 34 277 lexemes found in our whole corpus, we use some heuristics to remove non-lexical elements such as members of PUNCT, SYM, symbolic numerals, abbreviations, and certainly derived forms such as multiplicative numerals (e. g. *quinquies* 'five times'), or forms expressing a Degree (e. g. *ocior* 'swifter', with the *-ior-* comparative affix but no base form with the sole *oc-* stem). We end up focusing on 27 728 lexemes, from which we filter out all those recorded by WFL as the outcome of derivational processes by prefixation, suffixation or compounding, obtaining thus 19 076 candidates. Since WFL cannot cover all words, we manually inspect a random sample of 100 candidates to evaluate its precision,[36] assessing it at 42%. Among unknown, derived words we have *aleo* 'gamester', *mentalis* 'mental', or the compound *tricubitus* 'three cubits long'. Applying this value, we guess that the underived (not necessarily uninflectable) lexemes are ca. 8011, or 23.4% of the total (including noise). Among them, ADJ *turpis* 'ugly' (stem *turp-*) or NOUN *furnus* 'oven' (stem *furn-*). This means that, roughly speaking, almost three quarters of the Latin lexicon are derived from some more basic roots.

ality only continuous stems. Discontinuous stems (cf. Bonami and Beniamine, 2021) would only slightly change the picture: so, for example, for *facio* 'to do, make' we would obtain *f_c*, and then we could posit that its simplest continuous realisation *fac*, the imperative 'do', is a free morph.

[32] On the topic of predictability in Latin paradigms, even if for verbs, we refer to (Pellegrini, 2023).

[33] We notice that, at least etymologically, the stem variation *rump-/rup-* could be explained as the presence of still another affix, an imperfective nasal infix; cf. (Beekes, 2011, §12.1.5).

[34] The complete list is in the files aclitica_la.tsv, and with further data in aclitica_pos_la.tsv, in the repository.

[35] For example, the ADJ *ocior* 'swifter' not marked for the usual InflClass=IndEurX of comparatives, or *merito* 'being deserved = deservedly' not marked as a participial form of *mereo* 'to deserve' and instead annotated as an uninflectable ADV; on similar "contextual annotations" cf. (Cecchini, 2024, §5.1.1).

[36] The sample is in the file underived_random100_la_analysis.tsv in the repository.

This leads us to the perhaps obvious conclusion that the prototypical Latin non-clitic word is inflected, and in most cases includes derivational affixes of some kind, even when it is not inflectable (e. g. *nēquam*, with negative particle *ne-*, or *quandō*, with an ablative ending *-ō*). In MH's terms, we expect a Latin morph to be always accompanied by required and/or unrequired affixes, with free morphs occurring, but just as paradigmatic vagaries or not fully lexical elements (interjections).

6.1 Foreign words

We notice that, at this stage of analysis, we are ignoring the relevant category of terms of non-Latin, "foreign", origin, which would otherwise feature prominently among uninflectable words; most of these, not labelled for `InflClass`, are names, e. g. *dāuīd* 'David', but can also be common terms, e. g. *rabbi* 'teacher' (both from Classical Hebrew). However, such words cannot be really considered part of the Latin system as much as they are the remnant of a "failed" morphological integration process into Latin, possibly due to lack of analogy with other inflected words: we do see other foreign names being adapted, e. g. *iōannēs* 'John' (from Classical Hebrew through Ancient Greek) following the so-called third declension (here the inflectional affix is *-ēs*). Uncertainties in their treatment are seen also from the frequent choice of not annotating them in the treebanks (part of speech `X`, no features, `flat` relation), especially when in a non-Latin script (essentially, Greek). This category is thus problematic, because it is not clear when a word ceases to "belong" to one language and starts being part of another. To our ends, though, this is to some extent irrelevant: the uninflectability of these terms is a reaction to a different grammatical system, and not an internal evolution of Latin, at least not synchronically.

7 A good definition (for Latin)?

Trying to identify the morphological objects of MH's definition (see Table 2) for Latin has brought to light some issues.

The most immediate issue is that distinguishing a specific type of word for free morphs, as opposed to roots or clitics possibly appearing with affixes, is totally irrelevant for Latin, and actually for any language like Latin, that is, a language which tends to positively mark through inflectional morphology every term "denoting an action, an object or a property", independently from its semantic complexity (so, be it autosemantic or synsemantic), and where a good part of the lexicon, even uninflectable items, shows transparent extensive derivational processes. Only a fragmentary part of Latin forms can be identified as free morphs, but this does not really bring any insight into the nature of the Latin word. In fact, we get the impression that the inclusion of free morphs in MH's definition might be motivated, on the one hand, by the need to give a place to non completely lexical, difficult to tackle elements such as interjections, and, on the other hand, might be driven by the apparent focus on languages with different, if any, inflectional paradigms than Latin. In particular, we would like to stress the fundamental difference that incurs between a language which always marks some kind of grammatical category such as `Case`, `Number`, and/or `Gender` on the members of entire word classes, as is Latin, and other languages which only mark them in some circumstances and then possibly "additively", as it seems to be the case for English.

Even if we are not considering it in this study, we notice that MH's definition of compound will be directly affected by this parameter: if the members of some word class always occur with required affixes, we cannot expect them not to have some affix even when combining with other non-affixal morphs, and this is the case in Latin with the so-called linking vowel *-i-* (see e. g. Oniga, 1992; Brucale, 2012), so that we have to treat all such cases as instances of multiple, and not single, words. Conversely, we cannot expect such a thing as a linking vowel in languages with no required affixes, and so there we will treat most, if not all, morph combinations as single-word compounds. This, in our opinion, will then create an asymmetry in how and how many compounds are identified as single or multiple words in each language, which is ultimately based on a bias with respect to their respective inflectional types; this in turn, in our opinion, seems to run counter to the establishment of a definition which should apply universally with the same criteria, and which instead would take double standards, as it were. So we argue that it is not so much about deciding to always split forms or always lump them together, but that two cases like Latin *agricultor* 'farmer', from *ager* 'field' and *cultor* 'tender' with linking *-i-*, and *flowerpot*,

cited in (Haspelmath, 2023), should both be treated as either one single word or two words, but not as two words the former (Latin) and one word the latter (English), as would now be the consequence of a direct implementation of MH's definition.

Another major issue is that, while the distinction between roots and clitics is partly retraceable in UD's part-of-speech system, so morphosyntactically grounded, and partly reflected in phonological and morphological features, it still seems to not be really related to the presence or absence or affixes (for example, if we admit that AUX is clitic, then clitics can be inflected; also, some grammaticalised forms transparently bear derivational affixes), or to the status of boundness. In MH's scheme, what is common to all three basic word types is non-selectivity, and this seems to be more relevant: words could then be tentatively defined as *nuclei* of morphs and affixes which do not always syntactically combine with elements of the same kind. However, on the one hand both boundness and selectivity are not made explicitly relative to linear order or syntactic dependencies in MH's definition; and on the other hand, if we suppose the latter criterion, we can find counterexamples. For example, leaving further detailed investigation about this aspect to future work, we notice that a class such as ADP, clitic almost by definition, at least in Latin, always selects a nominal head to depend on, meaning a NOUN proper, a verbal form acting as a nominal (VerbForm=Vnoun), such as an infinitive (cf. Cecchini, 2021), or a modifier, which in Latin follows nominal morphology. In this sense an ADP is selective, but we cannot reasonably treat it as an affix, and this because it is not necessarily adjacent to the nominal in question, while conversely AUX is a much stronger candidate for being an affix given its tendency to stay close to its head. Unfortunately, the semantic criterion discriminating roots form other morphs does not come to help here in pinpointing what makes all these elements words, as has been observed in Section 5.1, and should probably be one of the first criteria to be discarded.

8 Conclusion

The conclusion of this preliminary study, at least for what concerns Latin, is that, unfortunately, MH's definition of word is less clear-cut than at first glance when it comes to putting it into practice, and it makes distinctions which do not appear relevant for the Latin system, and we dare to say also for typologically similar languages under the inflectional aspect, as could be Ancient Greek, Russian, and others.[37] In our opinion, some criteria of analysis, such as contentfulness, are already captured by UD's system in an even more systematical approach, and they are anyway (partially) orthogonal to the morphosyntactic ones that we deem have to be at the core of any definition of word, and that are also used in MH's definition. We see a trace, though, that deserves being pursued, and that is just based on selectivity and rigidity of linear order: words, especially in a morphosyntactic context as UD's formalism, can then be those blocks, with a meaningful nucleus and a possible contour of affixes, which can be moved around or separated by other similar blocks inside any possible free form. In this way, a multipartite classification into morphological objects as shown in Table 2, which seems to be more relevant at a language-specific level (e. g. more for English than for Latin), would be a consequence of a more general, and in our eyes useful and polished, definition. A similar definition is probably not new, but would still need to be applied uniformly throughout UD; for Latin, for example, it would have consequences on how to treat preverbs, i. e. verbal prefixes formally identical to adpositions and usually considered to be part of the stem (as seen in Section 6 for *ērumpo* 'to break out' with respect to *rumpo* 'to break'), and would probably go against some of the current traditional practices of annotation (so, for preverbs, it would mean splitting them from their bases). This issue, together with those about univerbation and compounding in Latin, are a material for future work which will be based on this first investigation on the very nature of wordhood in Latin. Beyond Latin, we hope that the scripts for data analysis that we make available with this study will help gather more data on this topic also for other languages, in order to gain a clearer picture of where the definition of word should be headed in a universal framework. At the same time, we are convinced that the particular case of Latin has helped putting

[37] We are well aware that this notion of similarity is impressionistic and that instruments need to be developed to make it quantifiable: this is by the way one of the aims making part of the development of a "tongueprint" for the ERASMOS project, at least for what concerns Latin and Ancient Greek.

a debated and crucial definition under a new perspective.

Acknowledgments

This work has been performed in the scope and with the support of the ERASMOS ERC-funded project (grant n. 101116087) at KU Leuven. Views and opinions expressed are those of the author only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

References

2024. LASLA/conllup.

J. N. Adams. 1976. A typological approach to Latin word order. *Indogermanische Forschungen (1976)*, 81:70–99.

David Bamman and Gregory Crane. 2011. The ancient Greek and Latin dependency treebanks. In *Language Technology for Cultural Heritage*, Theory and Applications of Natural Language Processing, pages 79–98, Berlin/Heidelberg, Germany. Springer.

Robert Stephen Paul Beekes. 2011. *Comparative Indo-European Linguistics*, second edition. Number 172 in Not in series. John Benjamins, Amsterdam, the Netherlands; Philadelphia, PA, USA.

Olivier Bonami and Sacha Beniamine. 2021. *Leaving the stem by itself*, pages 81–98. Number 353 in Current Issues in Linguistic Theory. John Benjamins, Amsterdam, the Netherlands.

Luisa Brucale. 2012. Latin compounds. *Probus – International Journal of Latin and Romance Linguistics*, 24(1 – On Romance Compounds):93–117.

Flavio Massimiliano Cecchini. 2021. *Formae reformandae*: for a reorganisation of verb form annotation in Universal Dependencies illustrated by the specific case of Latin. In *Proceedings of the Fifth Workshop on Universal Dependencies (*UDW*, SyntaxFest 2021)*, pages 1–15, Sofia, Bulgaria. The Association for Computational Linguistics (ACL).

Flavio Massimiliano Cecchini. 2024. Let's Do It Orderly: A Proposal for a Better Taxonomy of Adverbs in Universal Dependencies, and Beyond. *The Prague Bulletin of Mathematical Linguistics*, (121):15–65.

Flavio Massimiliano Cecchini, Timo Korkiakangas, and Marco Passarotti. 2020a. A New Latin Treebank for Universal Dependencies: Charters between Ancient Latin and Romance Languages. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation (LREC 2020)*, pages 933–942, Marseille, France. European Language Resources Association (ELRA).

Flavio Massimiliano Cecchini, Marco Passarotti, Paola Marongiu, and Daniel Zeman. 2018. Challenges in Converting the Index Thomisticus Treebank into Universal Dependencies. In *Proceedings of the Second Workshop on Universal Dependencies (*UDW *2018)*, pages 27–36, Brussels, Belgium. The Association for Computational Linguistics (ACL).

Flavio Massimiliano Cecchini, Rachele Sprugnoli, Giovanni Moretti, and Marco Passarotti. 2020b. UDante: First Steps Towards the Universal Dependencies Treebank of Dante's Latin Works. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020, Bologna, Italy, March 1–3 2021)*, pages 99–105, Turin, Italy. Associazione italiana di linguistica computazionale (AILC), aAccademia University Press.

James Clackson and Geoffrey Horrocks. 2007. *The Blackwell History of the Latin Language*. Blackwell Publishing, Malden, MA, USA.

András Cser. 2020. *The Phonology of Classical Latin*. Number 52 in Publications of the Philological Society. Wiley-Blackwell, Hoboken, NJ, USA.

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

Michiel de Vaan. 2008. *Etymological Dictionary of Latin and the other Italic Languages*. Number 7 in Leiden Indo-European Etymological Dictionary Series. Brill, Leiden, Netherlands; Boston, MA, USA.

Hanne Martine Eckhoff, Kristin Bech, Gerlof Bouma, Kristine Eide, Dag Haug, Odd Einar Haugen, and Marius Jøhndal. 2018. The PROIEL treebank family: a standard for early attestations of Indo-European languages. *Language Resources and Evaluation*, 52(1):29–65.

Federica Gamba and Flavio Massimiliano Cecchini. 2024. Sabellicus.

Federica Gamba, Abishek Stephen, and Zdeněk Žabokrtský. 2024. Universal Feature-based Morphological Trees. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (*MWE-UD*) @* LREC-COLING *2024*, pages 125–137, Turin, Italy. ELRA and ICCL.

Federica Gamba and Daniel Zeman. 2023a. Latin Morphology through the Centuries: Ensuring Consistency for Better Language Processing. In *Proceedings of the Ancient Language Processing Workshop associated with ehe 14th International Conference on Recent Advances in Natural Language Processing (*RANLP *2023)*, pages 59–67, Varna, Bulgaria. Incoma.

Federica Gamba and Daniel Zeman. 2023b. Universalising Latin Universal Dependencies: a harmonisation of Latin treebanks in UD. In *Proceedings of the Sixth Workshop on Universal Dependencies*

(UDW, GURT/SyntaxFest 2023), pages 7–16, Washington, D. C., USA. Association for Computational Linguistics (ACL).

Martin Haspelmath. 2017. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica*, 51(s1000 (Jubilee Issue: 50 Years Folia Linguistica)):31–80.

Martin Haspelmath. 2020. The morph as a minimal linguistic form. *Morphology*, 30:117–134.

Martin Haspelmath. 2021. Towards standardization of morphosyntactic terminology for general linguistics. In *Linguistic Categories, Language Description and Linguistic Typology*, number 132 in Typological Studies in Language, pages 35–58, Amsterdam, the Netherlands. John Benjamins.

Martin Haspelmath. 2023. Defining the word. *WORD*, 69(3):283–297.

Martin Haspelmath. 2025. Are "words" important for grammatical dependency analysis? handout. 3rd UniDive general meeting (January 29th, Budapest, Hungary).

Dag Trygve Truslew Haug and Marius Jøhndal. 2008. Creating a Parallel Treebank of the Old Indo-European Bible Translations. In *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pages 27–34, Marrakesh, Morocco. European Language Resources Association (ELRA).

Gerd Vanja Maria Haverling. 2000. *On Sco-Verbs, Prefixes and Semantic Functions*. Number LXIV in Studia Graeca et Latina Gothoburgensia. Department of Languages and Literatures, University of Gothenburg, Gothenburg, Sweden.

Federica Iurescia, Federica Gamba, Flavio Massimiliano Cecchini, Francesco Mambrini, Giovanni Moretti, Marco Passarotti, and Paolo Ruffolo. 2024. UD_Latin-CIRCSE.

Timo Korkiakangas and Marco Passarotti. 2011. Challenges in Annotating Medieval Latin Charters. *Journal for Language Technology and Computational Linguistics*, 26(2):103–114.

Christian Lehmann. 2020. Univerbation. *Folia Linguistica*, 54(s41-s1-Historica):205–252.

Eleonora Litta and Marco Passarotti. 2019. (When) inflection needs derivation: a word formation lexicon for Latin. In Nigel Holmes, Marijke Ottink, Josine Schrickx, and Maria Selig, editors, *Lemmata Linguistica Latina*, volume I: Words and Sounds, pages 224–239. De Gruyter, Berlin, Boston. Interrogable online at http://wfl.marginalia.it/.

Renato Oniga. 1992. Compounding in Latin. *Rivista di Linguistica (currently: Italian Journal of Linguistics)*, 4(1 – The Morphology of Compounding):97–116.

Marco Passarotti. 2019. The Project of the Index Thomisticus Treebank. In Monica Berti, editor, *Digital Classical Philology*, number 10 in Age of Access? Grundfragen der Informationsgesellschaft, pages 299–320. De Gruyter Saur, Berlin, Germany; Boston, MA, USA.

Matteo Pellegrini. 2023. *Paradigm Structure and Predictability in Latin Inflection*. Number 6 in Studies in Morphology. Springer, Cham, Switzerland.

Matteo Pellegrini. 2024. LatInfLexi version 2.0.1. https://zenodo.org/records/14438647.

Matteo Pellegrini and Marco Passarotti. 2018. LatInfLexi: an Inflected Lexicon of Latin Verbs. In *Proceedings of the Fifth Italian Conference on Computational Linguistics CLiC-it 2018*, pages 324–329, Turin, Italy. Accademia University Press.

Harm Pinkster. 2015. *The Oxford Latin Syntax*, volume 1. Oxford University Press, Oxford, UK.

Harm Pinkster. 2021. *The Oxford Latin Syntax*, volume 2. Oxford University Press, Oxford, UK.

Agata Savary, Daniel Zeman, Verginica Barbu Mititelu, Anabela Barreiro, Olesea Caftanatov, Marie-Catherine de Marneffe, Kaja Dobrovoljc, Gülşen Eryiğit, Voula Giouli, Bruno Guillaume, Stella Markantonatou, Nurit Melnik, Joakim Nivre, Atul Kr. Ojha, Carlos Ramisch, Abigail Walsh, Beata Wójtowicz, and Alina Wróblewska. 2024. UniDive: A COST Action on Universality, Diversity and Idiosyncrasy in Language Technology. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Underresourced Languages* @ LREC - COLING *2024*, pages 372–382, Turin, Italy. ELRA and ICCL.

Olga Spevak. 2010. *Constituent Order in Classical Latin Prose*. Number 117 in Studies in Language Companion Series. John Benjamin, Amsterdam, the Netherlands.

Leon Stassen. 2003. *Intransitive Predication*. Oxford Studies in Typology and Linguistic Theory. Oxford University Press, Oxford, UK.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Arofat Akhundjanova, Furkan Akkurt, Gabrielė Aleksandravičiūtė, Ika Alfina, Avner Algom, Khalid Alnajjar, Chiara Alzetta, Erik Andersen, Matthew Andrews, and 633 others. 2024. Universal Dependencies 2.15. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

A Distributions of free forms

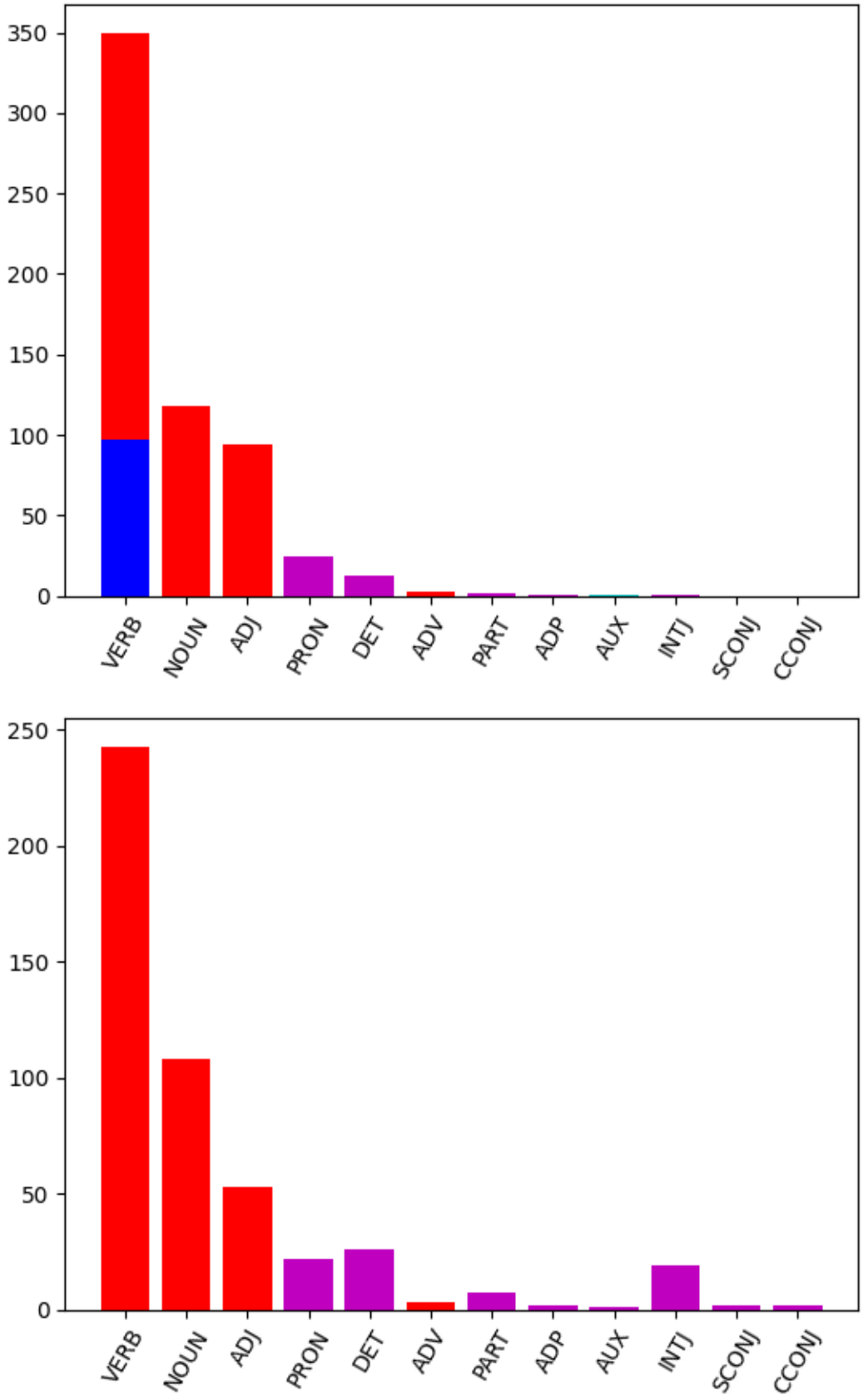

Figure 1: Distribution of parts of speech across our extracted free (above) and single (below) forms. Red is used for autosemantic and magenta for synsemantic classes; blue and cyan give the respective amount of one-word forms among the clausal free forms we define.