UD Segmentation Survey

Introduction

This survey, carried out as part of the UniDive COST Action CA21167, aims at collecting information on segmentation conventions in different UD treebanks, specifically on what constitutes a “word”. There are general guidelines on this, but as each language is different, we expect some divergence in how a word is defined in each UD treebank, some documented in the language-specific guidelines, some not.

As decided at the UniDive 1st general meeting, we will use Haspelmath’s (2023) definition of the word as a point of reference. We quote it here:

“A word is (i) a free morph, or (ii) a clitic, or (iii) a root or a compound possibly augmented by nonrequired affixes and augmented by required affixes if there are any.”

For the definitions of *lexeme*, *free form*, *morph*, *clitic*, *affix*, *root*, *compound*, and *required affix*, see the paper or this overview that we prepared.

Here are some examples of words, adapted from Haspelmath:

Type	**Examples**
Free morph	EN *nice*, EN *work*, EN *now*, EN *ouch*
Clitic	EN *the*, EN *to*, EN *’s*
Root	EN *tree*
Root plus required affixes	IT *alber-o*, IT *alber-i*, DE *geb-en*, DE *geb-t*
Root plus nonrequired affixes	EN *nice-r*, EN *go-ing*, EN *re-work*, EN *re-place-ment-s*
Root plus nonrequired affixes plus required affixes	DE *be-geb-en*, DE *be-geb-t*
Compound	EN *flower-pot*, EN *wind-shield*, EN *dog-sit*
Compound plus required affixes	DE *bau-spar-en*
Compound plus nonrequired affixes	EN *flower-pot-s*

Compound plus nonrequired affixes plus required affixes	DE *bau-spar-end-e*

Note that compounds are here defined in a quite restricted way (Haspelmath 2023, p. 287–288).

We appreciate your time filing in this survey.

Dan Zeman & Kilian Evang
Task 2.1 Co-leaders

Questions

Question 1: Who is filling out this form?

Flavio Massimiliano Cecchini, post-doctoral research fellow at CIRCSE (Università Cattolica del Sacro Cuore of Milan)

Question 2: Which language (or dialect) is this form being filled out for?

Latin (of many diachronical and style varieties)

Question 3: Some languages have more than one UD treebank, not necessarily all with the same segmentation conventions. If conventions differ, we ask you to fill out this survey separately for differing treebanks. Which treebank(s) is this form being filled out for?

ITTB, LLCT, Udante; indirectly also Perseus

Note: the first three treebanks are managed together by thesame group (CIRCSE), but due to different origins, might vary for minor details with regard to segmentation. However, the actively pursued aim is to make them converge. The Perseus treebank has historically had a different management, but is now also being made converge to the state of the art of the other three treebanks.

All following forms are present in at least one of the Latin treebanks; those which are not, but are shown for the sake of discussion, are marked with a *!* sign.

Question 4: Do *free morphs* occur as words[1] in the treebank(s)? If so, give a few examples. If not, say why (e.g., the language might not have them, or segmentation conventions for this treebank may not treat them as words).

Yes. For Latin, this is a minoritary case and we can actually list all occurrences (for indeclinable elements) and/or types we have in our data:

- **NOUN**
  - indeclinable: *fas* 'divine dictate'
  - forms in inflected paradigms: *tempus* 'time' (nom./acc. sg.), *iecur* 'liver' (nom./acc. sg.)
    - possibly also (see comments): *aer* 'air' (nom. sg.), *uir* 'man' (nom. sg.), *sol* 'sun' (nom. sg.), *manū* 'hand' (abl. sg.) < *manus*, *rē* 'thing' (abl. sg.) < *rēs*
  - foreign terms: *cenith* 'zenith' (Arabic), *cherub* (Hebrew), *ioseph* 'Joseph' (Hebrew)
- **ADJ**
  - *uetus* 'old' (n. nom./acc. sg.)
  - possibly forms in inflected paradigms (see comments): *celer* 'swift'
- **VERB**
  - forms in inflected paradigms: *fac* 'do (imperative 2 sg.)' < *facio* 'to do'
- **PRON**
  - forms in inflected paradigms: *tu* 'you' (nom. sg.)
- **DET**
  - indeclinable: *tot*, *quot* 'that many'
- **NUM**
  - indeclinable: *centum* '100', *decem* '10', *nouem* '9', *quattuor* '4', *quinque* '5', *septem* '7', *sex* '6'
- **ADV**
  - indeclinable: *palam* 'publicly'
    - Note: strictly speaking, *-am* could be considered a required, though crystallised, inflectional affix
- **INTJ**
  - indeclinable: *eheu* 'ah!'
  - foreign terms: *alleluia* (Hebrew, lit. 'praise god')

Notes:

1. the label SYM is not present in the treebanks;
2. DETs have to be considered free in Latin as they can appear independently, possibly with implicit/unspecified heads;
3. all other parts of speech not listed here (ADP, AUX, CCONJ, PART, SCONJ) have to be considered clitics in Haspelmath's sense (see next point);
4. ADVs are as of yet rather undefined, and so it is not predictable if an ADV is a free or a bound element. Further, they are problematic in the current annotation because, as it

1By *word*, we mean word in the UD sense, where a word can potentially be a part of a multi-word token (MWT).

stands, some forms in paradigmatic variation, e.g. ADV *nouiter* ‘fiercely’ vs ADJ *nouus* ‘new’, are treated disjointly: this means that, in such an annotation, nouiter superficially appears as a free morph, even if it is transparently analysable as *nou-iter* (and so would fall under point 7). There is a core of ‘true adverbs’, though, with no synchronically motivated derivation, such as the cited *palam*, or also *mox* ‘soon’ (with no recognisable affix).

Indeclinable elements, i.e. elements which represent the only form in the paradigm expected for their part of speech, are few and far between (and we might consider their appearance with or without unrequired affixes a random fact); those which also systematically coincide with a single morph are mostly represented by cardinal numbers, or by foreign terms which are not part of Latin’s system and have not been integrated into it (but some, like *sion* ‘hill by Jerusalem’ possess a double treatment).

In a Latin word we expect in general some required affix for inflection (see points 7, 9, 11, 13). However, we also find some systematic cases of forms in “regular” paradigms coinciding with a single morph, namey in the so-called IV (“u”) and V (“e”) nominal declensions in the ablative singular, and in other nominal declensions with nominatives in *-r* and *-l*. Actually here, given the fact that these occurrences seems to be explainable with regular and systematic phonetic phenomena such as *-r+s > -r and -l+s > -l*, one might venture to say that there is still a required inflectional affix, so we do not really have free morphs. The only candidates left would be neuter nouns of the III declension like *tempus*, which are in fact bare morphs (they also stay the same in the accusative).

Finally, a few imperative forms of verbs use the bare root, which otherwise always requires affixes: in general, there are no indeclinable VERBs/AUXs in Latin (at most defective paradigms).

Question 5: Do *clitics* occur as words in the treebank(s)? If so, give a few examples. If not, say why (e.g., the language might not have them, or segmentation conventions for this treebank may treat them as part of the word they attach to).

Yes. Traditionally, “clitic” in Latin is used for elements which are univerbated with another word, but in the sense of Haspelmath’s definition all members of the following parts of speech are regularly to be considered clitics:

- conjunctions, so **CCONJ** and **SCONJ**: *et*, *si*, *quam*...
  - *que* ‘and’ and *ue* ‘or’ are notable in that they are usually univerbated with the preceeding word, which is the last one of a co-ordinated phrase (but: !*SPQR* = ***S****enatus* ***P****opulus****q****ue* ***R****omanus* ‘the Senate and People of Rome ’): they are analysed in UD by means of multiword tokens
- adpositions, so **ADP**, mostly prepositions: *ad*, *cum*, *per*...
  - as for conjunctions, only a few are univerbated, e.g. *me****cum*** ‘with me’, *qua****tenus*** ‘to

what point'. Here we notice different treatments: while in the former case they are treated as parts of multiword tokens, in the latter this does not (always) happen (see comment below).

- discoursive and similar particles, usually **PART** (but also, somewhat inconsistently, **ADV**), very often appearing in fixed second ("Wackernagelian") position: *nam*, *ne, ...*
- the forms of the auxiliary **AUX** *sum*, and possibly also of *habeo* and *eo* when they act as auxiliaries
  - sometimes, when occurring after the main, lexical element, a form of *sum* is univerbated with it, but then always treated by means of multiword tokens
- it is questionable whether pronouns, so **PRON**, are bound rather than free forms, but probably the latter holds.

We observe a pattern here with regard to univerbation and/or wordhood: the breaking point for treating clitics as words in Latin seems to be their position with respect to the rest of the phrase. The expected, i.e. most frequent behaviour is when they appear at the left boundary of a phrase or internally (as discoursive particles with fixed second position): then there can be other elements between them and the head, and they are treated independently. Instead, when a clitic appears at the right boundary of a phrase, and so very often next to the phrase's head, very often it is written univerbated, and, if only functional elements are involved (also depending on the frequence of this behaviour), not even analysed independently: so, we observe cases such as the aforementioned *quatenus* tagged as ADV, and similarly *siquidem* (SCONJ *si* 'if, when' + PART *quidem* '(discoursive particle)'), *siquis* (SCONJ *si* 'if, when' + PRON *quis* 'that/which one'), and others.

A particular case is *quem+ad+modum* (DET *qui* 'that' + ADP *ad* 'at' + NOUN *modus* 'measure'), which in the treebanks appears both as a univerbated SCONJ and glossable as ‘'as for instance, or written and analysed individually (*quem ad modum*, though possibly featuring a lemma *quemadmodum*), acting as an oblique noun phrase. The same happens in other soruces also for a similar frase !*quam+ob+rem*, glossable as 'wherefore'.

Question 6: Do *roots* occur as words in the treebank(s)? If so, give a few examples. If not, say why (e.g., the language might not have them, or segmentation conventions for this treebank may not treat them as tokens).

Yes. The free morphs in point 4 belonging to lexical parts of speech (NOUN, ADJ, VERB) are also roots, and the same considerations apply.

They are in fact all roots (apart from INTJ) if Haspelmath's definition is relaxed a bit, in that "contentful" is also meant to extend to what in UD are functional parts of speech, so PRON, DET, AUX, while ADV stays excluded (again, this is a problematic class). As of now, a strict interpretation of the definition seems to point only to lexical elements (in UD's sense); further, *fas* 'divine law' or *tempus* 'time' are abstract concepts rather than "objects", so probably not "roots" (but still free morphs), and the same should hold for *cenith*.

Question 7: Do *roots plus required affixes* occur as words in the treebank(s)? If so, give a few examples. If not, say why (e.g., the language might not have them, or segmentation conventions for this treebank may not treat them as tokens).

The appearance of a root with required affix(es) is probably the expected, most common occurrence in Latin (cf. also point 9). As discussed under point 4, some apparent free roots might in fact be considered to contain a required inflectional affix.

In the following, the root is underlined and **bold** morphs are required affixes:

- **NOUN**
  - *lup*-***um*** ‘wolf (acc. sg.)’ < *lup*- ‘wolf’ (case-number affix)
  - *niu*-***es*** ‘snows (nom./acc.. pl.)’ < *niu*- ‘snow’ (case-number affix); nom. sg. *nix* (= *niu*- + -*s*)
    - possibly also *uir* (< *uir*+*s*) ‘man’ (nom. sg.), *sol* (< *sol*+*s*) ‘sun’ (nom. sg.)
- **VERB**
  - *fac*-***ieba***-***m*** ‘I was doing’ (act. ind. impf. past 1st sg.) < *fac*- ‘to do’ (TAM-voice-person-number affixes)
  - *fac*-***tur***-***um*** ‘[which is] going to be made’ (prosp. part., n. nom./acc. sg.) < *fac*- ‘to do’ (marker for syntactic role with [`VerbForm`], here adjectival [i.e. participle], plus case-gender-number affix)
  - *uul*-***t*** ‘wants’ (act. ind. impf. pres. 3rd sg.) < *uel*- ‘to want’ (TAM-voice-person-number affixe)
- **ADJ**
  - *long*-***arum*** ‘long (gen. f. pl.)’ < *long*- ‘long’ (case-gender-number affix)
  - *frug*-***i*** ‘useful’ (indecl.) < *frug*- ‘fruit’ (crystallised dat. sg. marker)
  - ... and also an adverbial form like *nou*-***iter*** ‘newly’ < *nou*- ‘new’ (not reflected in the current annotation style, see point 1)
- **ADP**
  - *sec*-***und***-***um*** ‘according to’ < *sequ*- ‘to follow’ (participial marker + gender-case-number affix), lit. ‘it going to follow’

Other similar combinations do not involve “roots” in the roughly corresponding UD sense of lexical elements with concrete meaning (see point 6).

Question 8: Do *roots plus nonrequired affixes* occur as words in the treebank(s)? If so, give a few examples. If not, say why (e.g., the language might not have them, or segmentation conventions for this treebank may not treat them as tokens).

Similar considerations as for point 4 hold: the presence or absence of a nonrequired affix on a free morph can be considered a random fact (and probably restricted to neuter nouns), and

indeclinable elements belonging here can be listed completely. Appropriate examples for VERBs seems to be nonexistent, for ADJs rare or marginal (or nonexistent if we take into accout the observation in point 4). Apart from PRONs, other non-lexical parts of speech do not seem attested here neither.

In the following, the root is underlined and **bold** morphs are the non required affixes:

- **NOUN**
  - *anim*-***al*** ‘animal’ (nom./Acc. sg.) < *anim*- ‘soul’ (nominaliser/adjectiviser?)
  - *lu*-***men*** ‘light’ (nom./acc. sg.) < *luc*- ‘to light’ (nominaliser)
  - ***in***-*star* ‘likeness’ (indecl.) < st- ‘to stand’ (lexical aspect prefix)
    - but *-ar* could be connected to the infinitive affix, or similar
  - possibly (but see point 4):
    - *ac*-***tio*** ‘action (nom. sg.)’ < *ag*- ‘to do’ (nominaliser)
    - ***con***-*sul* ‘consul’ (nom. sg.) < *sul*- ‘to take’ (lexical aspect prefix)
    - ***im***-*per*-***ator*** ‘emperor’ (nom. sg.) < *par*- ‘to provide’ (lexical aspect prefix, deverbal nominaliser)
  - Note: ***ne***-*fas* ‘impious deed’ (indecl.) < fas ‘divine law’ (negative particle) cannot be counted here, as the meaning of the base is abstract and so it is not a root according to Haspelmath
- **VERB**
  - there are apparently no cases where there is not also a required inflectional affix.
- **ADJ**
  - *long*-***ius*** ‘longer’ *(n. nom./acc. sg. cmp.)* > *long*- ‘long’ (degree)
  - probably none other (see point 4), but candidates could be:
    - !***per***-*celer* ‘very quick’ (nom. m. sg.) < *celer*- ‘swift’
      - NB: a variant of *celer*-*rim*-*us*
    - possibly (see point 4) *long*-***ior*** ‘longer’ (nom. f/m sg., cmp.) > *long*- ‘long’ (degree)
  - possibly indeclinable ***ne***-*quam* ‘worthless’ < *aequ*- ‘even (adj.)’ (negative particle) can be counted if the original *-am* case-number suffix is considered to have crystallised, else this would fall under point 9
- **PRON**
  - ***ni***-*hil* ‘nothing’ (nom./acc. sg.) < *hil*- ‘thread’ (negative particle)

Question 9: Do *roots plus nonrequired affixes plus required affixes* occur as words in the treebank(s)? If so, give a few examples. If not, say why (e.g., the language might not have them, or segmentation conventions for this treebank may not treat them as tokens).

Yes, and as commented in point 7 this is possibly the most common occurrence. Only a small selection of the myriad of possible examples follows.

In the following, the root is underlined and **bold** morphs are the non required affixes:

- **NOUN**
  - *ac*-***tion***-*is* ‘action (gen. sg.)’ < *ag*- ‘to do’ (nominaliser + case-number suffix)
  - *riu*-***ul***-*is* ‘rivulet’ (dat./abl. pl.) < *riu*- ‘stream’ (diminitiviser + case-number suffix)
  - ***con***-*sul*-*um* ‘consuls’ (gen. pl.) < *sul*- ‘to take’ (lexical aspect prefix + case-number affix)
  - ***im***-*per*-***ator***-*i* ‘emperor’ (dat. sg.) < *par*- ‘to provide’ (lexical aspect prefix + deverbal nominaliser, case-number affix)
  - *lu*-***min***-*ibus* ‘lights’ (dat./abl. pl.) < *luc*- ‘to light’ (nominaliser + case-number affix)
- **VERB**
  - ***in***-*sul*-***t***-*are* ‘to leap upon; to taunt’ (act. inf.) > *sal*- ‘to leap’ (lexical aspect prefix + iterative/frequentive suffix + marker for syntactic role [`VerbForm`], here nominal [i.e. infinitive])
  - ***ab***-*dic*-*amus* ‘we deny’ (act. ind. impf. pres. 1st pl.) > *dic*- ‘to say’ (lexical aspect prefix + TAM-person-number-voice affix)
  - ***ob***-*liu*-*isc*-*atur* ‘that he/she/it forget’ (act. subj. inch. pres. 3rd sg.) > *liu*- ‘livid; dark’ (lexical aspect prefix + TAM-person-number-voice affixes)
- **ADJ**
  - ***prae***-*clar*-*os* ‘very bright (m. acc. pl.) < *clar*- ‘bright’ (degree prefix + gender-case-number suffix)
  - *tim*-***id***-*us* ‘fearful’ (m. nom. sg.) < *tim*- ‘to fear’ (adjectiviser + gender-case-number suffix)
  - *leu*-***issim***-*um* ‘very soft’ (n. nom./acc. sg) < *leu*- ‘light’ (degree + case-gender-number suffix)

Question 10: Do *compounds* occur as words in the treebank(s)? If so, give a few examples. If not, say why (e.g., the language might not have them, or segmentation conventions for this treebank may not treat them as tokens).

Applying Haspelmath’s (2023) definition, there is probably no freely appearing compound in Latin. We notice that there is at least always either one of the following elements:

- a “linking element” between the two roots, usually *-i-*, which should probably be treated as a nonrequired affix;
- some required inflectional affix (this appears to be a strict requirement in Latin).

There are cases like !*dulc*-*acid*-*us* ‘bitter-sweet’ where it seems that the root *dulc*- ‘sweet’ has no linking element, but since this regularly happens for all first roots when the second root begins with a vowel, we should probably count such occurrences together with those showing an explicit linking element (this is a similar issue as that discussed under point 4 for “absorbed” inflectional affixes). In fact, we see it for the same elements in the complementary context, e.g. ! *dulc*-***i***-*flu*-*us* ‘sweetly flowing’. And sometimes the linking element appears also before a vowel,

as in !*acut*-***i***-*angul*-*um* ‘acute angle’.

The only possible occurrences of Haspelmathian compounds coudl be univerbations of numerals, if the notion of “root” is relaxed somewhat to include the abstract notion of “cardinal quantity” as a “property”. Then we observe

- *quin*-*decim* ‘15’ < *quinque* ‘5’ + *decem* ‘10’
- *si*-*decim* ‘16’ < *sex* ‘6’ + *decem* ‘10’
- possibly *du*-*centum* ‘200’ < *du*- ‘2’ + *centum* ‘100’
    - but we observe inflected *ducenta*, *ducentae*, *ducentis*..., as opposed to indeclinable *centum*

However, sometimes some of these combinations are also treated as separate, individual words, e.g. *uiginti quinque* ‘20 5’.

Only stretching the notion of root to also include DET, we might have

- **ADV**
    - *ho*-*die* ‘today’ < *ho*- ‘this’ + *die*- ‘day’
        - this originates however from a phrase in the ablative case

In general, elements belonging to functional parts of speech as a result of compoundig (in Haspelmath’s sense) appear to be rare to nonexistent.

In the following, only combinations of roots with demonstrably no linking element will be considered. Compounds of foreign words might exist, but then they should probably be treated as single elements (and so “free morphs”) in Latin.

In any case, all similar combinations of words, with or without (non)required affixes, are uniformly treated as individual words, with no MWT analysis, in Latin treebanks (more under point 14), following orthographic conventions. We note that at least in UDante the feature `Compound=Yes` is used for those combinations also found in thee Word Formation Latin lexicon (Litta & Passarotti 2019).

Question 11: Do *compounds plus required affixes* occur as words in the treebank(s)? If so, give a few examples. If not, say why (e.g., the language might not have them, or segmentation conventions for this treebank may not treat them as tokens).

Yes. As noted in point 10, inflectional requied affixes might actually be a strict requirement for combinations of roots in Latin.

Note: Similar forms are relatively sparse in data. With the help of Word Formation Latin (Litta & Passarotti 2019), 352 examples of combinations can be found, but many are not compounds in Haspelmath’s sense (i.e. they are not made up of roots), and most include a linking element

anyway. So, in the following, many examples do not actually appear in the treebanks, or appear only in PROIEL, but are shown here as a small case study.

In the following, the root is underlined and **bold** morphs are required affixes:

- **NOUN**
  - !*for*-*cep*-***s*** ‘pincers’ (nom. sg.) < *for*- ‘hot’ + *cap*- ‘to seize’ and case-number affix
  - !*puer*-*per*-***a*** ‘women who has given birth’ (nom. sg.) < *puer*- ‘male child’ + *par*- ‘to bear’ and case-number affix
  - !*iu*-*glan*-***s*** ‘walnut (tree)’ (nom. sg.) < *iou*- ‘Jupiter’ + *gland*- ‘acorn’ and case-number affix
  - !*au*-*spic*-***ibus*** ‘soothsayers’ (dat./abl. pl.) < *au(i)*- ‘bird’ + *spic*- ‘to look’ and case-number affix
  - *hos*-*pit*-***em*** ‘host’ (acc. sg.) < *host*- ‘stranger’ + *pot*- ‘able’ and case-number affix
    - this compound is so ancient that it can barely be analysed as such synchronically (see point 15)
- **ADJ**
  - *prin*-*cip*-***ibus*** ‘foremost [so, princes]’ (dat./abl. pl.) < *prim*- ‘first’ + *cap*- ‘to seize’ and case-number suffix
  - *parti*-*cip*-***es*** ‘partaking’ (nom./acc. pl.) < *parti*- ‘part’ + *cap*- ‘to seize’ and case-number-suffix
    - this of course only if the *-i* in *parti*- is taken as part of the root, which is debatable
  - *nau*-*frag*-***orum*** ‘shipwrecked’ (m./n. gen. pl.) < *nau*- ‘ship’ + *frang*- ‘to break’ and gender-case-number suffix
- **VERB**
  - *nun*-*cup*-***atur*** ‘is called by name’ (pass. ind. impf. pres. 3rd sg.) < *nom(en)*- ‘name’ + *cap*- ‘to seize’ and TAM-voice-person-number suffix
  - *iu*-*dic*-***amus*** ‘we judge’ (act. ind. impf. pres. 1st pl.) < *ius*- ‘law’ + *dic*- ‘to say’ and TAM-voice-person-number suffix
    - but in both latter cases it might be argued that ‘name’ and ‘law’ are abstract entities
  - *re*-*fer*-***t*** ‘it befits’ (act. ind. impf. pres. 3rd sg.) < *rē*- ‘thing’ + *fer*- ‘to bear’ and TAM-voice-person-number suffix
    - *rē* is interpreted as the ablative of *res*, coinciding in form with the bare root
  - *uen*-*d*-***eb***-***ant*** ‘they were selling’ (act. ind. impf. past 3rd pl.) < *uen*- ‘sale’ + *d*- ‘to give’ and TAM-voice-person-number suffix
  - *cre*-*d*-***it***-***is*** ‘trusted’ (pass. perf. participle, dat./abl. pl.) < *cord*- ‘heart’ + *d*- ‘to give’ and participial marker and case-number affix
    - the two latter compounds are so ancient that they can barely be analysed as such synchronically (see point 15)
- **CCONJ**
  - *sci*-*lic*-***et*** ‘that is’ < *sci*- ‘to know’ + *lic*- ‘to be available’ and TAM-voice-person-number suffix

Question 12: Do *compounds plus nonrequired affixes* occur as words in the treebank(s)? If so, give a few examples. If not, say why (e.g., the language might not have them, or segmentation conventions for this treebank may not treat them as tokens).

Apparently, there is no one that does not also contain the linking element, or does not bear a required inflectional affix, or where a possible nonrequired affix does not appear between roots. This goes hand in hand with the observations under points 4, 6, 8, 10. Such combinations do not seem to be structurally possible in Latin, or at least their occurrences are random facts, exceedingly rare and/or synchronically unproductive.

Question 13: Do *compounds plus nonrequired affixes plus required affixes* occur as words in the treebank(s)? If so, give a few examples. If not, say why (e.g., the language might not have them, or segmentation conventions for this treebank may not treat them as tokens).

Yes. Siilar schemata as under point 9 repeat also for combinations of roots. These "parasynthetic" constructions are always treated as individual words in the treebanks. They are probably less frequent and we often see the same combinations as in point 11 reappear.

In the following, the root is underlined and **bold** morphs are the non required affixes:

- **NOUN**
  - *nau*-*frag*-***i***-*o* 'shipwreck' (dat./abl. sg.) < *nau*- 'ship' + *frang*- 'to break' and nominaliser and case-number suffix
    - if we consider *-i-* a formative element and not part of the inflectional suffix (cf. *naufragorum*, point 11)
  - !*su*-*ově*-*taur*-***il***-***i***-*a* 'a sacrifice consisting of a swine, a sheep, and a bull' (nom./acc. pl.) < *su*- 'pig' + *oui*- 'sheep' + *taur*- 'bull' and adjectiviser and nominaliser and case-number affix
    - combining three roots
- **ADJ**
  - *prin*-*cip*-***al***-*is* 'original' (dat./abl. pl.) < *prim*- 'first' + *cap*- 'to seize' and adjectiviser and case-number suffix
- **VERB**
  - ***ad***-*iu*-*dic*-*au*-*it* 'has granted' (act. ind. perf. pres. 3rd sg.) < *ius*- 'law' + *dic*- 'to say' and lexcial aspect prefix and TAM-voice-person-number suffixes

Question 14: Can you think of any words in the treebank(s) that do not fall under any of the above categories (one example might be root+clitic where the clitic is not treated as a separate word)? Please give some examples and explain.

Yes, and there are in fact many types. The following is a tentative grouping.

1. As mentioned under point 5, there are quite many cases of combined roots and clitics, sometimes whole phrases, treated as single words (no MWT analysis) in Latin treebanks:
    1. **PRON *in*-*uic*-*em*** ‘each other’ (indecl.) < *uic*- ‘turn’ (adposition + crystallised inflectional affix), lit. ‘in turn [of smth/smb]’
        1. debatable if ‘turn’ is concrete or not, and hence a Haspelmathian root
    2. **ADV *de*-*nu*-*o*** ‘anew’ < *nou*- ‘new’ (adposition + crystallised inflectional affix)
    3. **ADV** *qu*-*em*-***ad***-*mod*-*um* ‘as for instance’ – see point 5
    4. **ADV** !***ob***-*iter* ‘passing along’ < *iter*- ‘way’ (adposition)

2. There are also many cases of elements composed of only clitics and/or functional elements:
    1. **ADV/SCONJ** *si*-*quidem* ‘if indeed’ < SCONJ *si* ‘if’ + PART *quidem* ‘(discoursive particle’)
    2. **ADV/CCONJ** *si*-*ue* ‘or if’ < SCONJ si ‘if’ + CCONJ -ue ‘or’ (univerbated clitic, see point 5)
    3. **PRON** *se*-***met***-*ips*-*am* ‘herself’ (f. acc. sg.) < PRON *s*- ‘self (reflexive)’ + emphatising affix + DET *ipse* ‘self (intensifier)’ and inflectional affix
    4. **DET** *quam*-*plur*-*es* ‘very many’ (m./f. nom./acc. pl.) < SCONJ *quam* ‘how (many)’ + DET *plur*- ‘many’ and inflectional affix

3. There are many combinations of roots with clitics and/or non-root elements (deictic adverbs, determiners...), among which numerals feature prominently:
    1. **ADV** *qu*-*a*-*re* ‘therefore’ < DET *qu*- ‘that’ + NOUN *re*- ‘thing’ and inflectional affixes
    2. **ADV** !*nu*-*di*-*us* ‘the present day’ < ADV *nu*- ‘now’ + *di*- ‘day’ and crysallised inflectional affix
    3. **NOUN** !*tri*-*um*-*uir*-***atu***-*s* ‘triumvirate, i.e. the rule by three appointed men’ (nom. sg.) < NUM *tri*- ‘three’ + NOUN *uir*- ‘male’ and inflectional affixes and nominaliser
        1. it could be argued over -*um*-, as it seems to coincide with the gen. pl. Suffix.
    4. **ADJ** !*sex*-*fasc*-***al***-*ium* ‘bearing six bundles’ (gen. pl.) < NUM *sex*- ‘six’ + NOUN *fasc*- ‘bundle’ and adjectiviser and inflectional affix

4. The bulk of “left-out” words is probably the one consisting of roots combined with a linking element, with and without affixes of any kinds (cf. point 10); examples are innumerable:
    1. **ADJ** *equ*-***i***-*noct*-***i***-***al***-*is* ‘pertaining to the equinox’ (m./f. nom. sg., gen. sg.) < ADJ *aequ*- ‘even’ + NOUN *noct*- ‘night’ and linking element and nominaliser and adjectiviser and inflectional affix

2. **NOUN** *agr-**i**-cul-**tor**-es* ‘farmer’ (nom./acc. pl.) < NOUN *agr-* ‘field’ + *col-* ‘to tend’ and linking element and nominaliser and inflectional affix
3. **ADJ** !*carn-**i**-uor-us* ‘carnivorous’ (m. nom. sg.) < NOUN *carn-* ‘flesh’ + VERB *uor-* ‘to swallow’ and linking element and inflectional affix
4. **VERB** *bell-**i**-ger-ant-es* ‘[those] waging war’ (act. impf. participle, m./f. nom./acc. pl.) < NOUN *bell-* ‘war’ + VERB *ger-* ‘to carry’ and linking element and syntactic role marker and inflectional affix
5. **ADV** !*ped-**e**-temp-tim* ‘step by step’ (indecl.) < NOUN *ped-* ‘foot’ + VERB *tend-* ‘to stretch’ and linking element (?) and adverbialiser

5. Finally, there are also whole phrases which are univerbated:
   1. **ADV** !*magn-oper-e* ‘greatly’ < ADJ *magno* ‘great’ (m./n. abl. sg.) + NOUN *opere* ‘work’ (abl. sg.) and inflectional affix
      - relation `det`, phrase in the ablative (so `obl`)
   2. **NOUN** *re-i-public-ae* ‘republic’ (dat. sg.) < NOUN *rei* ‘thing’ (gen. sg.)+ ADJ *publicae* ‘common’ (f. gen. sg.) and respective inflectional affixes
      - relation `amod`
   3. **NOUN** *patr-em-famili-as* ‘head of a family’ (acc. sg.) < NOUN *patrem* ‘father’ (acc. sg.) + NOUN *familias* ‘family’ (gen. sg.) and respective inflectional affixes
      - relation `nmod`
   4. **VERB** *uenun-d-ab-it* ‘he/she/it will sell’ (act. ind. impf. fut. 3rd sg.) < NOUN *uenum* ‘selling’ (acc. sg.) + VERB *do* ‘to give’ and inflectional affixes
      - relation `advcl`, a “secondary predication” of the object (lit. ‘to give [smth] as a selling’)
   5. **VERB** !*anim-**ad**-uert-iss-ent* ‘that they would have given attention’ (act. subj. perf. past. 3rd pl.) < NOUN *animum* ‘soul’ (acc. sg.) + VERB *aduerto* ‘to turn towards’ and inflectional affix
      - relation `obj`

Many other words, and some of the previous ones, would not be included in Haspelmath’s definition also because of the requirement of a root to be “concrete”.

There are also some other elements which are sometimes regarded as compounds (e.g. in WFL), but which in this framework should probably be better treated as bases with (nonrequired) affixes, e.g. PRON ***ali**-qu-is* ‘someone’ vs PRON *qu-is* ‘who’, DET *me-o-**pte*** ‘my very’ vs DET *me-o* ‘my’, PRON *ips-am-**met*** ‘her herself’ vs PRON *ips-am* ‘herself’ ADV, *quot-iens-**cumque*** ‘how often soever’ vs ADV *quot-iens* ‘how often’ vs DET *quot* ‘as many as’,DET *uter-**libet*** ‘whichsoever of the two’ vs DET *uter* ‘which of the two’, etc. We note that they are occasionally treated by means of MWT, e.g. *uos+met+ipsos* in UDante, so their analysis is not always uniform.

Some trends can be noticed:

- there is a strong tendency to put together, univerbated in orthography and analysed as

single words, functional words of various kinds which often occur together (case 2), even if they have syntactically different roles. An example is *siquidem*, where the almost fixed positions of the two elements make them frequent "neighbours". This custom might also be prompted by some of these elements being phoological clitics;

- This tendency also extends to roots (or, more in general, lexical elements in UD's sense) together with clitics or functional elements such as determiners (cases 1 and 3), and is most pronounced when the whole phrase appears in an oblique function: then it seems that the combination is seen almost as a "syntactic island" and is not further analysed, even when its structure its transparent and some of the elements also appear independently. Not all sources are uniform in the treatment of these combinations (so we have *quamobrem* vs *quam ob rem*, *uigintiquinque* vs *uiginti quinque*, etc.);
- combinations of roots with no inflection of the first member, or only the (prevalent) linking element, so compounds in the wider, traditional sense (case 4), are predominantly treated as single words with no further analysis; even more so in the case of "parasynthetic" constructions, especially with some overt derivational affix (e.g. *equinoctialis*). There can still be some oscillations and reinterpretations, as testified by a dictionary (Lewis & Short) remarking, under *agricultor*: "better separately, **agri cultor"** (NB: *agr-i* apparently coincides with the gen. sg. form of *ager*, but the vowel quantity is different);
- finally, there is a rather restricted groups of univerbated phrases (case 5) involving roots (or at least lexical elements) which, from source to source, can oscillate between a "monoword" teatment and a full syntactic analysis, such as *respublica*. Many of them seem to pertain to the legal semantic field (cf. Brucale 2012) and are distinguished by true compounds, since all elements involved show their inflectional affixes (and no linking elements), or, in the case of *animaduerto*, a verbal prefix (*ad-*) "unexpectedly" appears between the two roots. We can observe also an oscillation between forms such as *uenumdo* (inflectional affix), *uendo* (mere root juxtaposition), and even *uenum ... do* (not univerbated and possibly separated by other elements). In sum, these combinations cannot be regarded as compounds in any sense (cf. Brucale 2012, §2.2.2), and their treatment as single words in treebanks directly follows from spelling conventions.
    - *inuicem* (case 1.1) could also be grouped here, but (only in Late Latin?) we observe some occurrences where an analysis as oblique phrase does not seem to be viable anymore, e.g. UDante DVE-378 *ab invicem*, preceeded by an adposition.

Question 15: Do you have any additional relevant observations and remarks? Please list them here.

[addenda]

References

- Brucale, Luisa. "Latin compounds." Edited by Sergio Scalise and Francesca Masini. Probus – International Journal of Latin and Romance Linguistics (Berlin, Germany) 24, no. 1 – On Romance Compounds (2012): 93–117.
- Litta, Eleonora, and Marco Passarotti. "(When) inflection needs derivation: a word formation lexicon for Latin." In Words and Sounds, edited by Nigel Holmes, Marijke Ottink, Josine Schrickx, and Maria Selig, 224–239. Interrogable online at http://wfl.marginalia.it/. Berlin, Boston: De Gruyter, December 2019.