
UD Segmentation Survey 2.0

Preliminaries (1)

Which language is this form being filled out for? (1)

Type: (S/text-short)

Latin

Which treebank(s) is this form being filled out for, if any? If none, answer “N/A”. (2)

Type: (S/text-short)

UDante (mainly, but for Latin in general)

Who is filling out this form? (3)

Type: (S/text-short)

Flavio Massimiliano Cecchini

What is your email address? (4)

Type: (S/text-short)

flaviomassimiliano.cecchini@kuleuven.be

If this form is being filled out based on a response to the earlier UD segmentation survey, please give the name(s) of the original respondent(s). Otherwise, answer “N/A”. (5)

Type: (S/text-short)

Flavio Massimiliano Cecchini

Free morphs (2)

Give 3-10 examples of free morphs according to Haspelmath's definition in the target language that are also roots (denoting an action, an object, or a property). If there are none, answer "N/A".

Examples



Language	Example	Transliteration (if applicable)	Gloss (if applicable)	Type
en	work			action or object
en	John			object
en	he			object
en	nice			property
en	now			property
cs	pes		dog	object
cs	plyn		gas	object
bg	сив	siv	grey	property
de	sitz		sit.IMP.SING	action

(6)

Type: (./array-multi-flexi-text)

1 - L001	fac		do[IMP;2SG]	action
	Example - SQ001	Transliteration (if applicable) - SQ002	Gloss (if applicable) - SQ004	Type - SQ005
2 - L002	tempus		time[SG;NOM/ACC]	object
3 - L003	palam		publicly	property
4 - L004	manu		hand[SG;ABL]	object
5 - L005	fas		divine_dictate	object
6 - L006	uetus		old[N;SG;NOM/ACC]	property
7 - L007	i		go[IMP;2SG]	action
8 - L008	mox		soon	property
9 - L009				
10 - L010				

Comments (7)

Type: (T/text-long)

Uninflectable, non-foreign roots are extremely rare.

In some declensions, some inflected forms correspond to bare roots and no inflectional morph can be identified.

Neuter property/object roots of the third declension systematically appear as bare roots in the nominative/accusative case.

Bare root verbal forms are only observed for some verbs and only in the imperative.

Foreign terms left uninflected (e.g. cenith 'zenith', ierusalem 'Jerusalem') might be considered as roots, but they are not really integrated in Latin's system and are not considered here.

Give 3-10 examples of free morphs according to Haspelmath's definition in the target language that are *not* roots.

Examples

Language	Example	Transliteration (if applicable)	Gloss (if applicable)
en	ouch		
cs	sakra		damn
bg	леле	lele	wow
de	ja		yes

(8)

Type: (/array-multi-flexi-text)

1 - L001	tot		so_many
	Example - SQ001	Transliteration (if applicable) - SQ002	Gloss (if applicable) - SQ003
2 - L002	eheu		ah
3 - L003	centum		(one) hundred
4 - L004	ita		thus
5 - L005	non		not
6 - L006	tu		you[2SG;NOM]
7 - L007			
8 - L008			
9 - L009			
10 - L010			

Comments (9)

Type: (T/text-long)

There are only two uninflectable determiners, the demonstrative/relative couple tot/quot.

All cardinal numerals above 3 are uninflectable, with possible exceptions of some powers of 10.

Most forms of personal pronouns are probably analysable as bare roots.

Are free morphs according to Haspelmath always treated as words in the treebank? (10)

Type: (L/list-radio)

Unsure

AO03

If you answered “No” in the previous answer, give counterexamples. (11)

Type: (/array-multi-flexi-text)

1 - L001	totidem		so_many- same	just as many	idem is itself a determiner used adverbially in its neuter form
	Example - L001	Transliteratio n (if applicable) - L005	Gloss (if applicable) - L004	Translation (if applicable) - L003	Comment - L002
2 - L002	nonnullus		not-no_one	some	nullus is itself ullus 'any' + negating prefix n-
3 - L003	nondum		not-as	not yet	dum is not a free morph
4 - L004	uigintiquinque		20-5	twentyfive	Combination of free morphs
5 - L005					
6 - L006					
7 - L007					
8 - L008					
9 - L009					
10 - L010					

Comments (12)

Type: (T/text-long)

Combinations of "small (i.e. syncategorematic) words" are often unvierbated and (not) analysed as such.

Numerals are also often kept together as non-analysed entities.

I am confused about a point: given free morphs which are part of larger paradigms, what should it mean to "always treat them as words"?

Clitics (3)

Give 3-10 examples of clitics according to Haspelmath's definition in the target language. If there are none, answer "N/A".

Examples

Language	Example	Transliteration (if applicable)	Gloss (if applicable)	Translation (if applicable)
en	the			
en	to			
en	's			
en	small *and* cute			
cs	Sm-ál *se*.		laugh-PAST.MASC.SING *REFLEX*	He laughed.
es	*de* la escuel-a		of the.FEM.SING school-FEM.SING	of the school
de	*auf* der Brücke		on the.FEM.SING.DAT bridge.FEM.SING. DAT	on the bridge
bg	*Ще* ти кажа.	*Šte* ti kaža.	Will you.2PERS.SI NG.DAT tell.1PER S.SING.PERF.PRE S	I will tell you.

(13)

Type: (/array-multi-flexi-text)

1 - L001	et			and
	Example - SQ002	Transliteration (if applicable) - SQ005	Gloss (if applicable) - SQ004	Translation (if applicable) - SQ003
2 - L002	si			if
3 - L003	que			and
4 - L004	ad			to
5 - L005	cum			with
6 - L006	nam			therefore
7 - L007	ne		question	
8 - L008				
9 - L009				
10 - L010				

Comments (14)

Type: (T/text-long)

Clitics occurring at the right margin of a phrase are most often univerbated (e.g. -que, -cum), but still treated as individual syntactic words, at least with respect to content words (see below).

Are clitics according to Haspelmath always treated as words in the treebank? (15)

Type: (L/list-radio)

No

AO04

If you answered “No” in the previous answer, give counterexamples. (16)

Type: (/array-multi-flexi-text)

1 - L001	quatenus		which-F;SG;A BL-as_far_as	how far	qua often not recognised as part of the paradigm of qui
	Example - L001	Transliteratio n (if applicable) - L005	Gloss (if applicable) - L004	Translation (if applicable) - L003	Comment - L002
2 - L002	siquidem		if-indeed	if indeed	
3 - L003					
4 - L004					
5 - L005					
6 - L006					
7 - L007					
8 - L008					
9 - L009					
10 - L010					

Comments (17)

Type: (T/text-long)

Combinations where clitics appear on the right margin AND which involve function words are very often (and traditionally) not further analysed in their components, beyond being univerbated. So we have populusque 'and the people' analysed as two words, but siquidem very often as one. A part of the explanation is that, while left-margin clitics (e.g. ADP) can be separated from the head by other elements, this does not happen for right-margin clitics.

This usage can go as far as to involve entire phrases, mostly of "adverbial" character, e.g. quemadmodum = quem + ad + modum that-F/M;SG;ACC_to_measure-SG;ACC = 'as for instance', but here there is much more oscillation.

Roots (with affixes) (4)

Give 3-10 examples of *roots* denoting *actions* . Try to include both *bound roots* (if the language has them) and *free roots* (if the language has them). For *bound roots* , include required affixes as well so that you get a word, but separate them with hyphens. Include glosses in single quotes.

Examples

Language	Example	Transliteration (if applicable)	Gloss (if applicable)	Translation (if applicable)	Type
en	go				free
en	work				free
en	place				free
cs	*plyn*-out		flow-INF		bound
bg	*глед*-ам	gled-am	look-1.SG.PRE S		bound
de	lauf		run.IMP		free

(18)

Type: (/array-multi-flexi-text)

1 - L001	fac		do[2SG;IMP]	do	free
	Example - SQ001	Transliteration (if applicable) - SQ002	Gloss (if applicable) - SQ003	Translation (if applicable) - SQ005	Type - SQ004
2 - L002	*fac*-ieba-m		do-PST;IPFV- 1SG;ACT	I was doing	bound
3 - L003	*fac*-tur-um		do-PRSP;ACT -N;SG;NOM/AC C	[which is] going to be made	bound
4 - L004	*uul*-t		want-3SG;ACT	wants	bound
5 - L005					
6 - L006					
7 - L007					
8 - L008					
9 - L009					
10 - L010					

Comments (19)

Type: (T/text-long)

Only a few imperative forms correspond to roots in Latin. Otherwise, action roots can be regarded as being always bound and requiring conjugational affixes.

Give 3-10 examples of *roots* denoting *objects* . Try to include both *bound roots* (if the language has them) and *free roots* (if the language has them). For *bound roots* , include required affixes as well so that you get a word, but separate them with hyphens. Include glosses in single quotes.

Examples

Language	Example	Transliteration (if applicable)	Gloss (if applicable)	Translation (if applicable)	Type
en	tree				free
en	he				free
en	John				free
cs	plyn		gas		free
bg	*жен*-и	*žen*-i	*woman*-PL.F EM		bound
de	*Häus*-er		*house*-PL		bound

(20)

Type: (/array-multi-flexi-text)

1 - L001	fas			divine dictate	free
	Example - SQ001	Transliteration (if applicable) - SQ002	Gloss (if applicable) - SQ003	Translation (if applicable) - SQ005	Type - SQ004
2 - L002	*lup*-um		wolf-SG;ACC	wolf	bound
3 - L003	*niu*-es		snow- PL;NOM/ACC	snows	bound
4 - L004	tempus		time[SG;NOM/ ACC]	time	free
5 - L005	*tabul*-atim		plot-ADVZ	in rows	bound
6 - L006	*frug*-i		fruit-SG;DAT	worthy	bound
7 - L007					
8 - L008					
9 - L009					
10 - L010					

Comments (21)

Type: (T/text-long)

There are some forms in some paradigms which look like free morphs, like *honor*, but they can be explained with regular sound laws like $r+s > r$, and thus we can interpret them regularly as bound morphs with the inflectional affix, as the stems having different shapes. The prototypical Latin object root always requires inflectional affixes.

This is different for neuter stems in the accusative/nominative singular of so-called third declension, which appear systematically as bare morphs.

frugi is a transparent inflected form from *frux* which is traditionally counted as uninflectable adjective.

Adverbial forms like *tabulatim* are very often not seen as part of a paradigm.

Give 3-10 examples of *roots* denoting *properties* . Try to include both *bound roots* (if the language has them) and *free roots* (if the language has them). For *bound roots* , include required affixes as well so that you get a word, but separate them with hyphens. Include glosses in single quotes.

Examples



Language	Example	Transliteration (if applicable)	Gloss (if applicable)	Translation (if applicable)	Type
en	nice				free
cs	*plyn*-ul-y		flow-PART-MA SC.SING.NOM	fluent	bound
cs	ně-*jak*-y		INDEF-how-MA SC.SING.NOM	some, any	bound
bg	сив	siv	grey		free
de	zwei		two		free

(22)

Type: (/array-multi-flexi-text)

1 - L001	uetus		old[N;SG;NO M/ACC]	old	free
	Example - L001	Transliteratio n (if applicable) - L004	Gloss (if applicable) - L003	Translation (if applicable) - L002	Type - L005
2 - L002	*long*-arum		long- F;PL;GEN	(of) the long (ones)	bound
3 - L003	*celer*-es		swift-M/F;PL;N OM/ACC	swift (ones)	bound
4 - L004	*nou*-iter		new-ADVZ	newly	bound
5 - L005					
6 - L006					
7 - L007					
8 - L008					
9 - L009					
10 - L010					

Comments (24)

Type: (T/text-long)

Even more so than object roots, Latin property roots always require inflectional affixes. If not, they do not inflect (but there are only one/two like this).

The same notes for object roots hold here, e.g. for *celer*, *uetus*.

Adverbial forms like *nouiter* are very often not seen as part of a paradigm, or as showing a required affix.

Give 3-10 examples of roots with nonrequired affixes. (Required affixes are of course also allowed.) If there are none, answer "N/A".

Examples



Language	Example	Transliteration (if applicable)	Gloss (if applicable)	Translation (if applicable)	Comment
en	*tree*-s				plural -s not required
en	re-*place*-ment				neither re- nor -ment required
tr	*ev*-ler		house-PL		plural -ler not required
de	*lauf*-en		run-INF		infinitive -en not required
de	Ver-*flach*-ung		MUT-flat-NMLZ	flattening	ver- required after -ung is added

(23)

Type: (/array-multi-flexi-text)

1 - L001	*lu*-men		light-NMZ[SG; NOM/ACC]	lamp	Bare neuter form in the otherwise inflected paradigm
	Example - L001	Transliteration (if applicable) - L004	Gloss (if applicable) - L003	Translation (if applicable) - L002	Comment - L005
2 - L002	in-*star*		in-to_stand	appearance	Uninflectable
3 - L003	im-*per*-ator		in-to_provide-NMZ-SG;NOM	emperor	We can consider it as imperator-s with regular sound change r+s>r. Prefix in-related to action root, not to the noun
4 - L004	ne-*fas*		not-divine_dictate	impious deed	Uninflectable, as the base
5 - L005	ne-*quam*		not-even	worthless	Uninflectable; base might be

				crystallised case form
6 - L006	*ac*-tion-is	put_in_motion- NMLZ-SG;GEN	(of the) action	Inflected form of possibly bare actio
7 - L007	in-*sul*-t-are	in-to_jump-DI M-INF;IMPFV;A CT	to leap into (later: to taunt)	Non-required affixes are independent from each other and are related among others to Aktionsart; frequentative form, here rendered as diminutive
8 - L008	*leu*-issim-um	light-SUP-N;S G;NOM/ACC	very soft	Degree affix is not required
9 - L009	*lu*-min-ibus	light-NMLZ- PL;ABL/DAT	(from/to) lamps	Inflected form of lumen
10 - L010	*riu*-ul-is	stream-DIM- PL;ABL/DAT	(from/in the) rivulet	

Comments (25)

Type: (T/text-long)

It needs to be seen how many object/property roots do not require some nominaliser/adjectivaliser affix, whose non-requiredness would then become questionable, with respect to action roots.

Are roots with or without affixes always treated as words in the treebank? (26)

Type: (L/list-radio)

No

AO02

If you answered “No” in the previous answer, give counterexamples.

(27)

Type: (/array-multi-flexi-text)

1 - L001	quamobrem		that-F;SG;AC C-towards-thing- SG;ACC	for which reason	
	Example - L001	Transliteration (if applicable) - L005	Gloss (if applicable) - L004	Translation (if applicable) - L003	Comment - L002
2 - L002	reipublicae		thing-SG;DAT- public- F;SG;DAT	(to the) republic	lit. the common thing
3 - L003	iureconsulto		law-SG;ABL-t o_consider-PT CP;PFV;PASS- M;SG;DAT/ABL	to one (man) skilled in law	consulo is itself a combination of required and unrequired affixes
4 - L004	iurisiurandi		law-SG;GEN-I aw-PTCP;PRS P;PASS- N/M;SG;GEN	of the oath	
5 - L005	magnopere		great-N;SG;A BL/DAT-work- SG;ABL/DAT	greatly	lit. with great work; the suffix of the first

				element is phonologically regularly fused, and this is represented in writing
6 - L006	uenundabit	purchase-SG; ACC-to_give-P ST;IPFV-3SG;A CT	was selling	the accusative form uenum is phonologically adapted also in writing
7 - L007				
8 - L008				
9 - L009				
10 - L010				

Comments (28)

Type: (T/text-long)

For many of these cases, especially nominal combinations, the treatment can be arbitrary and depend completely on the possible graphical univerbation, which is not consistent.

These are not compounds because the single parts each follow an independent inflection.

Compounds (with affixes) (5)

Give 3-10 examples of compounds according to Haspelmath in the target language.

Examples

Language	Example	Transliteration (if applicable)	Gloss (if applicable)	Translation (if applicable)
en	flower-pot			
cs	straš-pytel		fear-bag	scaredy-cat
de	Auto-bahn		car-way	highway
bg	кандидат-студент	kandidat-student	candidate-student	

(29)

Type: (/array-multi-flexi-text)

1 - L001	*prin*-*cip*-*ibus		first-to_seize- PL;ABL/DAT	(from/to the) foremost (ones)	
	Example - L001	Transliteration (if applicable) - L005	Gloss (if applicable) - L004	Translation (if applicable) - L003	Comment - L002
2 - L002	*iu*-*glan*-*s		Jupiter-acorn- SG;NOM	walnut	lit. Jupiter's acorn
3 - L003	*nun*-*cup*-*atur		name-to_seize- -PRS;IPFV;3SG ;ACT	is called by name	Phonological assimilation of nom- (as in nomen) > nun-)
4 - L004	*sci*-*lic*-*et		to_know-to_be _available-PRS;	that is	

		IPFV;3SG;ACT	
5 - L005	*prin*-*cip*-al- is	first-to_seize- ADJZ- F/M;SG;NOM	(from/to the) principal (ones)
6 - L006	*nau*-*frag*-i- o	ship-to_break- NMLZ- SG;ABL/DAT	(from/to the) shipwreck
7 - L007	ad-*iu*-*dic*-a u-it	to-law-to_say- PFV- PST;3SG;ACT	has granted
8 - L008			
9 - L009			
10 - L010			

Comments (30)

Type: (T/text-long)

There is often at least a linking element (see examples below) which strictly speaking makes it not possible to speak about compounds in a strict Haspelmathian sense. I feel that this issue should be discussed, as the basic definition probably needs a refinement.

"Fused compounds" (i.e. no linking element) are in many cases barely recognisable as such anymore, as they have often undergone phonetic changes.

There are combination of bare non-root morphs, though, such as quin-decim 'fifteen'.

Also, the requirement of at least inflectional affixes seems to be even more severe for combinations of roots in Latin, even where there is apparently no linking element, as in dulc-acid-us, but these cases can be explained with regular sound laws.

Are such compounds always treated as words in the treebank? (31)

Type: (L/list-radio)

Yes

A001

If you answered “No” in the previous answer, give counterexamples. (32)

Type: (/array-multi-flexi-text)

1 - L001					
	Example - L001	Transliteration (if applicable) - L005	Gloss (if applicable) - L004	Translation (if applicable) - L003	Comment - L002
2 - L002					
3 - L003					
4 - L004					
5 - L005					
6 - L006					
7 - L007					
8 - L008					
9 - L009					
10 - L010					

Comments (33)

Type: (T/text-long)

Other UD words (6)

If the language has compounds (consisting of two or more roots, with affixes) with linking elements that are treated as words in the treebank, list examples here. Otherwise, answer “N/A”.

Language	Example	Transliteration (if applicable)	Gloss (if applicable)	Translation (if applicable)
de	Liebe-s-brief		love-GEN-letter	love letter
cs	ruk-o-pis		hand-LINK-write	manuscript
bg	син-е-ок	sin-e-ok	blue-LINK-eye	blue-eyed

(34)

Type: (/array-multi-flexi-text)

1 - L001	*agr*-i-*cul*-tor-es		field-LINK-to_tend -NMLZ- PL;NOM/ACC	farmers
	Example - L001	Transliteration (if applicable) - L005	Gloss (if applicable) - L004	Translation (if applicable) - L003
2 - L002	*equ*-i-*noct*-i-al-is		even-LINK-night- NMLZ-ADJZ- F/M;SG;NOM	pertaining to the equinox
3 - L003	*ped*-e-*temp*-tim		foot-LINK- to_stretch-ADVZ	step by step
4 - L004				
5 - L005				
6 - L006				
7 - L007				
8 - L008				
9 - L009				
10 - L010				

Comments (35)

Type: (T/text-long)

If the language has contractions that are split in the treebank, list examples here in the shown form. Otherwise, answer “N/A”.

Language	Example	Transliteration (if applicable)	Split form	Transliteration (if applicable)	Gloss (if applicable)
de	am		an dem		at the.MASC.SI NG.DAT
de	im		in dem		in the.MASC.SI NG.DAT
fr	au		à le		to the.MASC.SIN G

(36)

Type: (/array-multi-flexi-text)

1 - SQ001	scriptumst		scriptum est		to_write-PTCP ;PFV;PASS-N;S G;NOM/ACC-to-be-3SG;PRS
	Example - L001	Transliteration (if applicable) - L007	Split form - L006	Transliteration (if applicable) - L005	Gloss (if applicable) - L004
2 - SQ002	sin		si ne		if-not
3 - SQ003					
4 - SQ004					
5 - SQ005					

Comments (37)

Type: (T/text-long)

They are mostly very rare, if not non-existent in texts such as those of UDante. Contractions were part of the spoken language but are usually not represented in more formal texts, or edited.

When they do appear, mostly function words are involved.

If the language has any other forms that are treated as words in the treebank but are not words according to Haspelmath, list examples here. (38)

Type: (/array-multi-flexi-text)

1 - L001	hodie		this-M/N;SG;ABL/DAT-day-SG;ABL	today	Fused inflected determiner and noun, transparent
	Example - L001	Transliteration (if applicable) - L005	Gloss (if applicable) - L004	Translation (if applicable) - L003	Comment - L002
2 - L002	inuicem		in-turn-SG;ACC	each other	Has different syntactic distribution than expected by a noun phrase
3 - L003	denuo		from-new-M/N;SG;ABL;DAT	anew	
4 - L004	obiter		towards-way	passing along	
5 - L005	semetipsam		self[REFL;3SG;ABL/ACC]-EMPH-self-F;SG;ACC	her herself	Where -met can be seen as an affix, but is sometimes analysed separately
6 - L006	siue		if-or	or if	
7 - L007	quamplures		how_many-many-F/M;PL;NOM/ACC	very many	
8 - L008	sedecim		six-ten	sixteen	
9 - L009	sexfasculum		six-bundle-ADJZ-PL;GEN	(of the ones) wearing six bundles	
10 - L010					

Comments (39)

Type: (T/text-long)

Beyond the cases previously considered, these elements always include or consist only of syncategorematic elements. For them, there is traditionally a strong tendency to universion and analysis as a single word.

Additional remarks (7)

Are there any other remarks or observations that you would like to share? (40)

Type: (T/text-long)

Orthographic conventions are often a strong incentive to the treatment as individual words. The confusion between syntactic words and greater lexical units is also part of the problem. Function words are often treated differently than content words: there is a tendency to fuse the former together, while analysing the latter separately. A core/oblique distinction is also active: "adverbial" phrases are more liable to an unsegmented annotation, even when their structure is transparent. This is often done even when spelling conventions are not consistent in the sources.

Haspelmath's definition of compounds needs to be refined by taking into account the "obligatoriness of inflection" of a given language: a term as English flowerpot is not more a compound than Latin agricultor (with the linking element -i-), and they both should be analysed as multiwords, as it were. This is because English has no "expected inflection", and so we do not expect a linking element at all. This of course leads to a much restricter definition of what is a "true" (i.e. "fused") compound in languages such as English.