Word Segmentation in Universal Dependencies

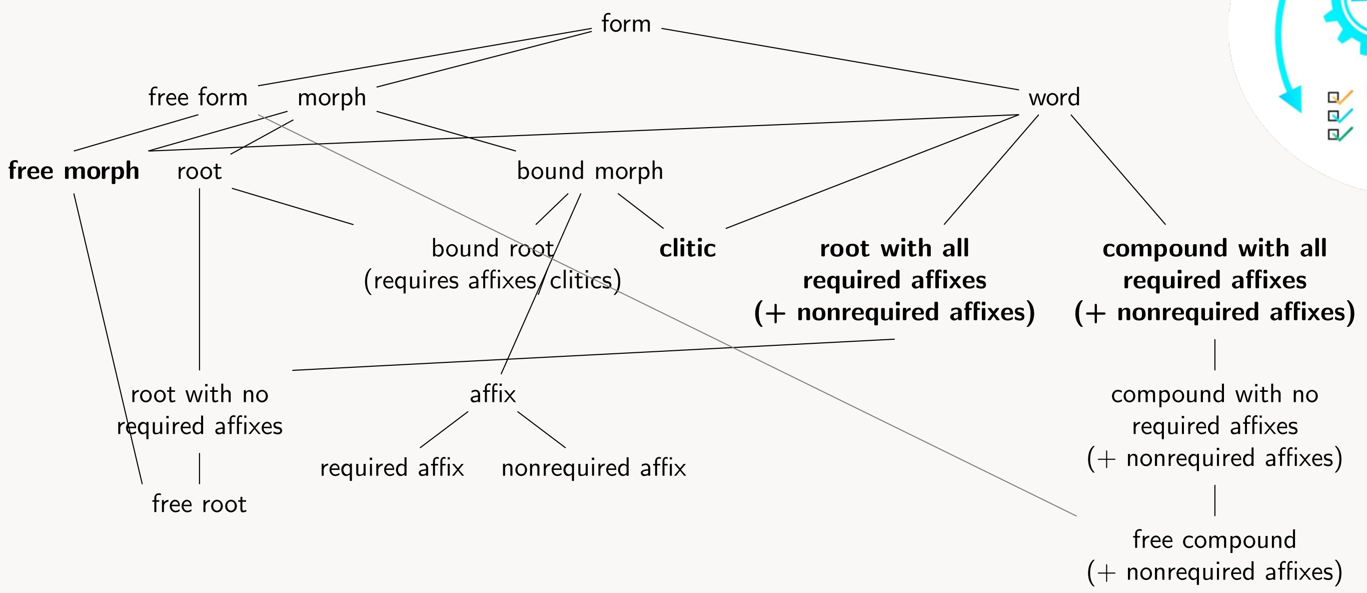
Introduction

- Notion of “word” central to UD but hard to define
- Are UD treebanks consistent wrt. this? Used Martin Haspelmath’s paper *Defining the Word* as a point of comparison

Tokens and Words in UD

es «¡Vámonos al mar!», exclamó Juan. (untokenized)
es « ¡ Vámonos al mar ! » , exclamó Juan . (tokens)
es « ¡ Vamos nos a el mar ! » , exclamó Juan . (words)

Haspelmath’s Terminology



form
free form
morph
word
free morph
root
bound morph
bound root
(requires affixes/clitics)
clitic
root with all required affixes (+ nonrequired affixes)
compound with all required affixes (+ nonrequired affixes)
root with no required affixes
affix
compound with no required affixes (+ nonrequired affixes)
required affix
nonrequired affix
free root
free compound (+ nonrequired affixes)

- nodes in bold are **words** in Haspelmath’s terminology

Free Morphs

en	nice (property)
en	work (action or object)
en	now (property)
cs	pes ‘dog’ (object)
en	ouch (not a root)

Roots (+Affixes)

cs	plyn ‘gas’ (free root)
tr	ev-ler ‘house-Plur’ (free root with affix)
it	alber-o ‘tree’ (bound root with required affix)
en	re-place-ment (non-required affixes)
cs	Josef-ov-ým ‘Josef-Poss-Ins’ (opt ⇒ req affix)

Clitics

en	**the** book
cs	Smál **se**. ‘He laughed.’
es	**de la** escuel-a ‘of the school’;
de	**auf der** Brücke ‘on the bridge’

Compounds (+Affixes)

en	flower-pot
de	Auto-bahn ‘highway’
cs	straš-pytel ‘scaredy-cat’
el	γεω-γραφ-ία / geô-graf-ía ‘geography’ (affix)

Non-Haspelmath Words in UD

de	Liebe-s-brief ‘love letter’
cs	ruk-o-pis ‘manuscript’
de	am ⇒ an dem ‘at the’, im ⇒ in dem ‘in the’
fr	au ⇒ à le ‘to the’

Survey

- 1.0: received responses for 43 languages; terminology challenging
- 2.0: in progress; as Google form, more structured collection of examples, attempt to elucidate terminology more (cf. diagram above), consider also non-UD languages

Main Areas for Harmonization

- demarcation of clitics (words) vs. affixes (non-words)
- compounds: always split?
- crosslinguistically applicable criterion for when to split contractions

Kilian Evang, Daniel Zeman
evang@hhu.de, zeman@ufal.mff.cuni.cz
Heinrich Heine University Düsseldorf, Charles University

UniDive
cost EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY
Funded by the European Union