UniDive WG2/WG1 Meeting Napoli

Defining the Word

Daniel Zeman, Kilian Evang, Petya Osenova (and Martin Haspelmath)

February 9, 2024

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

BY NC SA
unless otherwise stated

Why

- NLP assumes a word-like unit
- Difficult to define cross-linguistically
  - Especially for spoken data, unwritten languages, or some writing systems (Chinese)
  - Linguists have been arguing since (at least) 1930's, no satisfactory outcome
- Picking a definition (or not picking any) impacts all levels of annotation

Why

- NLP assumes a word-like unit
- Difficult to define cross-linguistically
  - Especially for spoken data, unwritten languages, or some writing systems (Chinese)
  - Linguists have been arguing since (at least) 1930's, no satisfactory outcome
- Picking a definition (or not picking any) impacts all levels of annotation

- If we can find a reasonable definition, we probably cannot force everyone to use it
  - Compatibility with legacy tools, resources etc.

Why

- NLP assumes a word-like unit
- Difficult to define cross-linguistically
    - Especially for spoken data, unwritten languages, or some writing systems (Chinese)
    - Linguists have been arguing since (at least) 1930's, no satisfactory outcome
- Picking a definition (or not picking any) impacts all levels of annotation

- If we can find a reasonable definition, we probably cannot force everyone to use it
    - Compatibility with legacy tools, resources etc.
- **BUT:**
    - Guidance for new languages that do not have a strong (or any) tradition
    - Or for tricky phenomena within a language
    - Or at least a principled method of documenting the deviations

How

Martin Haspelmath (2023). Defining the word. In *Word*, 69:3, pp. 283–297, Routledge.
https://doi.org/10.1080/00437956.2023.2237272

- A recent proposal
- Claims to have overcome some shortcomings of previous definitions
- We decided to test it on UD languages

- We organized a survey in summer – fall 2023
- Identified difficult points that are not addressed in the paper
- Going to do a second survey, better instructions

Prerequisities

Haspelmath's definition of *word* relies on a number of other concepts. Some of them he defines in a possibly surprising way.

- Free form
- Morph
  - Delimiting morphs
  - Recognizing "same" morphs
- Root
  - Recognizing roots
  - Classifying roots (object, property, action)
- Affix
- Clitic
- Compound

Free Form

A form (=non-empty sequence of phonemes) that can occur alone (at least as a response to a question).

- 🇬🇧 *He is going home.*
- 🇬🇧 *he*
- 🇬🇧 *he is*
- 🇬🇧 *he is going*
- 🇬🇧 *home*

- 🇨🇿 *pes* ‘dog’
- 🇬🇧 *the dog* (while *dog* itself is not a free form)
- 🇬🇧 *ouch!*

Morph

Smallest meaningful (lexical or grammatical) segment. Non-empty!
MH avoids the term morpheme.

- 🇬🇧 *dog-s* (lit. *dog-**Plur***)
- 🇨🇿 *ps-ů* (lit. *dog-**Plur.Gen***)
- 🇹🇷 *köpek-ler-in* (lit. *dog-**Plur**-**Gen***)

- 🇬🇧 *en-large-ment-s*
- 🇨🇿 *ob-děl-at* (lit. *PREF-do-**Inf***) ‘cultivate’

Morph

Smallest meaningful (lexical or grammatical) segment. Non-empty!
MH avoids the term morpheme.

- 🇬🇧 *dog-s* (lit. *dog-***Plur**)
- 🇨🇿 *ps-ů* (lit. *dog-***Plur.Gen**)
- 🇹🇷 *köpek-ler-in* (lit. *dog-***Plur-Gen**)

- 🇬🇧 *en-large-ment-s*
- 🇨🇿 *ob-děl-at* (lit. *PREF-do-***Inf**) 'cultivate'

- 🇩🇪 *zu-m* (lit. *to-***Def.Masc,Neut.Sing.Dat**) (= *zu dem* 'to the')
- 🇩🇪 *zu-r* (lit. *to-***Def.Fem.Sing.Dat**) (= *zu der* 'to the')
- 🇫🇷 ? *au* (= *à le* 'to the')
- 🇵🇹 ? *à* (= *a a* 'to the')

“Same” Morph

- Same or compatible meaning
  - Not the same: 🇬🇧 *I can swim* vs. *This can is empty*
  - Not the same: 🇬🇧 *I have two book-s* (plural) vs. *She book-s the flights* (third person singular present)

“Same” Morph

- Same or compatible meaning
  - Not the same: 🇬🇧 *I can swim* vs. *This can is empty*
  - Not the same: 🇬🇧 *I have two book-s* (plural) vs. *She book-s the flights* (third person singular present)
- Same form or just phonological changes
  - 🇬🇧 *-ed* [d] or [t] or [əd] (past tense)
  - 🇩🇪 *Träne-n* (lit. *teardrop-**Plur***) and *Burg-en* (lit. *castle-**Plur***) ... phonological variants
  - 🇩🇪 *Mutter* (lit. *mother.**Sing***) and *Mütter* (lit. *mother.**Plur***) ... not the same meaning
  - 🇩🇪 *Sohn* (lit. *son.**Sing***) and *Söhn-e* (lit. *son-**Plur***) ... phonological variants?
  - 🇨🇿 *pes* (lit. *dog.**Sing**.**Nom***) and *ps-a* (lit. *dog-**Sing**.**Acc***) ... phonological variants
  - 🇨🇿 *matk-a* (lit. *mother-**Sing**.**Nom***) and *matc-e* (lit. *mother-**Sing**.**Dat***)
  - 🇹🇷 *ev-ler* (lit. *house-**Plur***) and *araba-lar* (lit. *car-**Plur***)

Free Morph vs. Bound Morph

- Free morph
  - A morph that is a free form, i.e., it can occur alone.
  - 🇬🇧 *nice*
  - 🇬🇧 *ouch!*
  - 🇹🇷 *ev* ‘house’ … may have affixes (*ev-ler, ev-ler-de*) but they are not required
  - 🇨🇿 *dům* ‘house’ … may have affixes (*dom-y, dom-ech*) but they are not required

- Bound morph
  - A morph that always occurs in company of other morphs.
  - 🇬🇧 *house* … needs article *(a house, the house)* or plural suffix *(house-s)*
    - The articles and plural suffix are bound morphs as well
  - Adpositions
    - Although in some languages debatable (answer to a question about the preposition?)
  - 🇨🇿 *My se to-ho ne-boj-íme* ‘We are not afraid of this’

Root

- Contentful morph (denoting object, property, or action)
- Can occur in a free form without another contentful morph
  - This does not mean that the root itself is free

- Root categories matter (to distinguish affixes from clitics)
  - **object** / **property** / **action**
- Avoid POS categories such as NOUN / ADJ / VERB
  - 🇬🇧 Property root *large* as ADJ, but also in VERB *(enlarge)* or NOUN *(enlargement)*
- BUT:
  - Sometimes a root is not itself clearly in one category
    - 🇨🇿 *plyn* 'gas' vs. *plyn-out* 'flow'
  - Abstract nouns are not objects – are they properties?
  - But they behave the same way.
    - 🇨🇿 *matk-a, matk-y, matc-e* 'mother'
    - 🇨🇿 *lásk-a, lásk-y, lásc-e* 'love'

Root

- Pronouns – do they count as object roots?
    - Probably yes

- Numerals – do they count as property roots?
    - Probably yes

Affix

- Bound morph that accompanies roots of one category
- Not necessarily adjacent to the root – there can be other affixes (but not roots or clitics) in between
- Inflectional affixes, derivational/lexical affixes

- 🇹🇷 *ev-ler-de* (lit. *house-**Plur**-**Loc***) ‘in houses’
- 🇬🇧 *re-place-ment*
- 🇨🇿 *Josef-ov-ými* (lit. *Josef-**Poss**-**Plur.Ins***) ‘Josef’s’
- 🇪🇸 *geo-grafía* ‘geography’
- 🇺🇦 *Він см-іяв-ся (Vin sm-ijav-sja)* (lit. *He laugh-**Past**-**Refl***) ‘He laughed’

Clitic

- Bound morph that accompanies roots of different categories
- (This definition is different from what people may think a clitic is.)
- **Unclear:** What does it mean that a clitic accompanies a root? What if there are multiple roots in the free form where the clitic occurs?

- 🇨🇿 *b-át se* (lit. *fear-*_**Inf Reflex**_)
- 🇨🇿 *My se medvěd-a ne-boj-íme* ‘We are not afraid of a bear’
- 🇪🇸 *reanimar-se* ‘to revive oneself’
- 🇪🇸 *se ha pod-ido observ-ar* ‘it could have been observed’
- 🇪🇸 *se hab-ía comenz-ado a investig-ar* ‘one had begun to investigate’

Compound

- Two or more roots immediately adjacent (without linking material, i.e., no non-root morphs in between)
- The non-head root cannot be modified separately by a nominal or an adjective.

- 🇩🇪 *Auto-bahn* (lit. *car-way*) ‘highway’
- 🇬🇷 *γεω-γραφ-ία (geô-graf-ía)* ‘geography’
- 🇬🇧 *credit card*

- **Not compounds:**
- 🇩🇪 *Liebe-s-brief* (lit. *love*-**Gen**-*letter*) ‘love letter’
- 🇨🇿 *ruk-o-pis* (lit. *hand-o-write*) ‘manuscript’

Word

- Free form (may or may not be root)
- Root with all required affixes (and possibly with other affixes)
- Compound with all required affixes (and possibly with other affixes)
- Clitic

Martin Haspelmath (p.c.): *"My take at the moment is that dependencies should ultimately be notated at the morph level, and that the lexicon vs. syntax distinction is illusory (along the lines of Jackendoff)."*