Utrecht
University

WG 2.4: Canonical Form for MWEs

Jan Odijk

Budapest, 2025-01-28

Overview

- Definitions
- Principles
- Role of Lemma
- Lexicon requirements
- Annotation requirements

Definitions

- **MWE**: UniDive definition
- **Multiword expressions** (MWEs) are (continuous or discontinuous) sequences of words with the following compulsory properties:
  - They show some degree of orthographic, morphological, syntactic or semantic idiosyncrasy …
  - include a headword and at least one other syntactically related word. …
  - At least two components of such a word sequence have to be lexicalized.

Definitions

- **Canonical Form (CF):** a unique form that represents a set of forms
- **Example: lemma:** single form to represent all inflectional variants of a word
  - Which forms are considered inflectional variants?
  - Which form is selected as the lemma ?
  - ➔depends on the language, grammatical theory, part of speech, lexicographer

Definitions

- **CF for MWEs**: a single form that represents all variations of the MWE that are due to grammar (lexical variation requires a different CF)
  - this differs from other uses of this term in Parseme/UniDive)

Definitions

- **Examples:**
  - variations due different grammatical features (e.g. singular v. plural, present tense v. past tense, etc.)
  - different grammatical constructions: productive voice alternation, e.g. passivization; relativisation; verb placement rules; verb cluster formation; order variations; …
- **Grammatical variation must be made explicit for each language**
- **Working doc:** **https://docs.google.com/document/d/18UeTgOcSIVH0yw7gJnLqtRYPO07ZIBV40rWNWmdAfOc/edit?usp=sharing**

Principles

- Initial set of Guiding Principles
  - For determining the CF
  - For generalising from the word form in the CF to other word forms
- A CF has features to specify properties of its components:
  - Inflectability, Modifiability, Head, fixedness, ...
  - default values by the guiding principles, but can be overruled
  - Non-default values represented by annotations on the CF (annotated CF). Example

Role of Lemma

- A CF contains word forms and lemmas
- Lemma form as head implies full inflectability as default property
- Which form is selected as the lemma, is not relevant

Requirements (lexicon)

- Possibly multiple CFs for ‘one MWE’:
  - *Met zijn <vieren>* v. *met =zijn <vieren>*
  - *Lit. with his fours, `there’s four’*
  - *Optional arguments:*
    - *Iemand 0zal 0de gelegenheid geven <om ..>*
      Someone will the opportunity give to ..
    - *Iemand 0zal iemand 0de gelegenheid geven <om ..>*
      Someone will someone the opportunity give to ..

Requirements (Annotation)

- Occurrence of an MWE must have as annotations:
    - Parseme annotations: span, MWE type
    - But also:
        - Lexicon name, lexicon version
        - MWEId: ID of the lexicon entry
        - CF: canonical form of the MWE
        - Query (type) used, Matching node id, head position,
        - Automatic (software & version) or manual (annotator id), date/time
    - Standardise these annotation attributes

Thanks for you attention!

Grammatical Variants

- E.g. in Dutch: ***iemands hart*** *breken (*break sm’s heart)
  - <possessive np> *hart*, e.g. *mijn tantes hart* (my aunt's heart)
  - *het hart van* <np>, e.g. *het hart van de buurman* (*the heart of the neighbour*)
  - <np> *z'n/d'r/hun hart*, e.g. *haar vriend z'n hart* (*her friend his hear*t)
  - <possessive pronoun> *hart*, e.g. *mijn hart* (*my heart*)

Example CF

- CF: *Someone will take advantage of something*
  - *will*: not a component of the MWE (marking: 0)
  - *someone, something*: free variables
  - *take*: head, modifiable (default)
  - *take*: inflectable (head in lemma form, no marking)
  - *advantage*: modifiable (non default, marking: *)
- Annotated CF:
  - *Someone 0will take *advantage of something*