WG 1: Lexicon-Grammar Interface

Manfred Sailer Gyri S. Losnegaard

2nd General Parseme Meeting, Athens

March 11th, 2014

Schedule for WG 1 sessions

1. Session 9: Discussion
   - Classification of MWEs
   - MWEs in Computational Lexica

2. Session 10: Organization
   - WG-internal Communication
   - Plan till and for Haifa 2014

Overview

1. Session 9: Discussion
   - Classification of MWEs
   - MWEs in Computational Lexica
2. Session 10: Organization
   - WG-internal Communication
   - Plan till and for Haifa 2014

Classification of MWEs

- Basis: Sag et al. (2002), Baldwin and Kim (2010)
- General principles for the classification:
  - type of idiomaticity: lexical, syntactic, semantic, pragmatic, statistical
  - internal syntactic structure
  - degree of flexibility: fixed, semi-fixed, flexible, institutionalized MWEs
- Based on English, i.e. a language
  - with little inflection,
  - with little syntactic flexibility
- Problems:
  - individual examples
  - other languages
- Proposal

Individual example 1: Degree of flexibility

(1) einen Streit vom Zaun brechen (German)
a quarrel from the fence break 'start a fight'

- Ellsiepen (2005)
- institutionalized combination: *Streit* more frequent than other head nouns of the direct object
- fixed *vom Zaun*:
  - no *von dem*-alternation:

    (2) *einen Streit von dem Zaun brechen

  - modification attested, though not common (external modification?):

    (3) wurde... die Debatte... vom parteipolitischen Zaun
    was the debate from the party-political fence
    gebrochen (IDS)
    broken

Constituent parts of MWEs may differ with respect to their flexibility!

Individual example 2: Size of the MWE

- Restricted lexical variation:

  (4) lose one's mind/marbles

- Systematic variation:

  (5) let/take/put/. . . the cat out of the bag

- Predicative use:

  (6) a. She seems/looks a bit under the weather.
  b. With Caleb under the weather, Sports Director Darren Kinnard changed up Martin's Moments a bit . . . (www)

- Negation:

  (7) $\left\{ \begin{array}{c} \text{Pat didn't} \\ \text{Nobody could} \\ \text{*Pat could} \end{array} \right\}$ get a word in edgewise.

Individual examples 3: Additional MWE patterns

- phraseological patterns: *The X-er the Y-er*, ...
- binomials: *alive and well*
- N-idioms beside compounds: *piece of cake*, *Archilles' heel*
- additional VP idioms patterns: *barking up the wrong tree*, *call it a day*
- AP idioms: *fit as a fiddle*
- PP idioms: *under the weather*
- CP/S idioms: *You can say that again.*
- ...

(taken from en.wikipedia.org/wiki/List_of_English-language_idioms, en.wikipedia.org/wiki/Siamese_twins_%28linguistics%29)

Other languages 1: Similar test — different interpretation

- Fronting not the same in all languages:
  - English topicalization only for flexible idioms
  - German 'Vorfeld' also for semi-flexible idioms

(8) a. *The bucket Pat kicked. (no MWE reading)
b. Those strings Pat pulled to get Kim the job. (MWE reading)

(9) a. [Den Löffel] hat er abgegeben. (German)
the spoon has he given-away
('He has died.')
b. Aber [die Fäden] hat der Senior noch in der Hand.
but the strings has the senior still in the hand
('But the senior boss is still pulling the strings')

Other languages 2: Different test — what interpretation?

Slavic languages: perfective vs. imperfective aspect

(10) Vanja popleval/ *napleval v potolok. (Russian)
Vanja-nom perf-spit/ *perf-spit in ceiling
('Vanja frittered away time.')
projects.chass.utoronto.ca/fasl15/abstracts/BabyonyshevKavitskaya_ab#96.pdf

Proposal

- General ideas
- Concrete suggestions
- Example:
  www.lexical-resource-semantics.de/wiki/index.php/Parseme_WG1

General principles 1: Flexibility

- Flexibility should be motivated by morpho-syntactic criteria, not by semantic/pragmatic criteria.
- Flexibility determined for each subconstituent of an MWE.

(11)

einen	Streit	vom	Zaun	brechen	(German)
a	quarrel	from the	fence	break	('start a fight')

  - [einen Streit]: typical representative; other expressions of fight also possible
  - [vom Zaun]: semi-fixed

General principles 2: Degrees of flexibility

- Degrees: Fixed, semi-fixed, flexible, institutionalized
  - Fixed:
    no formal variation possible
  - Semi-fixed:
    only obligatory, meaningless formal variation possible (such as inflection, German V2, . . .)
  - Flexible:
    formal variation is possible that suggests that parts of an MWE contribute meaning (such as internal modification, clefting, . . .)
  - Institutionaliized:
    Full formal variation, but striking co-occurrence of MWE parts

General principles 2: Structure of the template

- Language-specific tests for flexibility for various syntactic categories and syntactic patterns
- Language-specific list of syntactic patterns attested for MWEs
- (Examples of MWEs with indication of their degree of flexibility and syntactic structure)
- Illustration: for each test. Ideally with naturally occurring examples, examples from the literature, also ungrammatical examples

Example template

Preliminary url:
www.lexical-resource-semantics.de/wiki/index.php/Parseme_WG1

Points of discussion:

- Guideline: Size of an MWE?
- Example: More syntactic patterns?
- Language-specific tests/phenomena?

Overview

1. Session 9: Discussion
   - Classification of MWEs
   - MWEs in Computational Lexica
2. Session 10: Organization
   - WG-internal Communication
   - Plan till and for Haifa 2014

MWE lexicons

WG1 objectives

- Better understanding of linguistic properties of MWEs, in particular at the lexical and syntactic level
- Enhancing the usability of MWE lexicons and valence dictionaries in parsing
- Paving the way towards interoperability of lexicons and the reduction of their production cost.

MWE lexicons

How can we achieve this?

MWE lexicons: construction and design

Central issues:

- What do we mean by MWE lexicon?
- What information to encode?
  - How much information about the possible variations of an MWE do we actually need to include in the lexicon in order to parse it?
  - Are there properties that should be represented in the lexicon although they are not directly relevant for parsing?
  - Do we actually need to put syntactically flexible MWEs in the lexicon?
- And how?
  - Representation frameworks
  - Analysis of the MWE
- Language and purpose specific considerations

Lexicon related issues: WG outcomes

We can:

- Make recommendations for MWE lexicon development.
- Identify and encourage best practices.

Lexicon related issues: WG outcomes

How to get there:

- Describe our own resources.
- Special volume on MWE lexicon construction and development.
- Develop an ISOcat taxonomy for MWE description?
- Other?

Special volume on MWE lexicon development

Proposal:

A (more or less) uniform presentation of lexical resources developed/used (etc.) by PARSEME members, each contribution having the same basic structure and addressing a few common issues.

For instance:

- a description of the project/research
- the structure of the lexicon
- the linguistic properties encoded
- a discussion of the reusability of the MWE lexicon
  - e.g.: is the format idiosyncratic to that resource, or have some kind of standard format been used?
  - could a similar resource be built for other languages?

Development of an ISOcat taxonomy for MWEs

Standardization task.

Other examples of standardizations efforts:

Universal Dependency Annotation for Multilingual Parsing (http://ryanmcd.com/papers/treebanksACL2013.pdf).

Lexicon related issues: WG outcomes

1st year outcomes

-
- mailing lists,
-
- website,
- detailed scientific program of each WG,
- internal share spaces,
- contrastive state-of-the-art surveys in all WGs,
- workshop proceedings,
- publications & technical documents,
- annual report.

Overview

1. Session 9: Discussion
   - Classification of MWEs
   - MWEs in Computational Lexica

2. Session 10: Organization
   - WG-internal Communication
   - Plan till and for Haifa 2014

WG1 now

Information on the Parseme page:

typo.uni-konstanz.de/parseme/index.php/2-general/51-wg-1-lexicon-grammar-interface

	September 2013	March 2014
Members:	31	47
Countries:	19	21
Female/Male:	17/15	24/23
ESR/non-ESR:	12/19	21/26

BSCW folder

- https://bscw.server.uni-frankfurt.de/
- All WG1 members have received invitation (accepted: 25)
- Password-protected area for sharing:
  - Your own files (papers, drafts, data collections)
  - Links to publications
  - Files for project-internal use
- Ideal for sharing chapter/article drafts for joint edited publications
- Further options: forum, blog, calendar
- But: no wiki

Wiki

- Proposal: WG-internal wiki hosted in Frankfurt
- Password-protected
- Content:
  - MWE templates for individual languages
  - Overview sheets for MWE resources
- Wiki url: wiki.studiumdigitale.uni-frankfurt.de/FB10_Parseme

Communication

- Mailinglists
  - parseme-all
  - parseme-wg1
- Inside the BSCW folder? (Forum or blog?)

Overview

1. Session 9: Discussion
   - Classification of MWEs
   - MWEs in Computational Lexica

2. Session 10: Organization
   - WG-internal Communication
   - Plan till and for Haifa 2014

Working plan for year 1 (till March 2014)

- Comparative study of MWEs in the parseme languages:
  - Definition of an MWE-template
  - Filling in the template for some example languages
- Collection of lexical resources of MWEs (links, descriptions, tools, ...)

Tasks till Haifa

- Communication: Trying out the platform(s)
- MWE template:
  - Authors for MWE template for individual languages
  - Work on MWE pages for individual languages
- Lexical resources
  - Create overview sheets for MWE resources — till May 15
  - Publication:
    - ★ After May 15: Evaluate the submitted resources
    - ★ Look for publication place, topic
    - ★ Discussion via the mailing list

WG 1 sessions in Haifa

- Joint WG1 meetings or subgroups (1. Linguistic properties; 2. Lexical formalisms)? **No!, but parallel sessions in Haifa**
- Poster session
- Linguistic properties (1 session):
  - Problems with the MWE template
  - Planning a joint publication (volume with OALI, *Phraseologie und Parömiologie*, . . . on MWE classification in various languages)
- Lexical formalism (1 session):
  - Articles for the lexicon development publication
  - If a suitable person is available: Invite a key representatives of standardization initiatives (MWE modelling)

Preliminary working plan for year 2 from Warsaw

- Internal and external reviewing of collaborative publication
- Enrichment of lexical resources on MWEs
- With WG 2: Example integration of MWE representations into grammar implementations
- With WG 4: Discussing WG1's classification criteria with tree bank annotation

Adjusted working plan for year 2

- Communication: BSCW folder, Wiki
- Linguistic properties:
  - MWE classification template for x languages (on the wiki)
  - Preparation of a joined publication, based on the classification templates
    (Haifa: concept; Malta: discussion of first versions)
- Lexical formalisms:
  - Information sheets for lexical resources (on the wiki)
  - Preparation of a joined publication, based on the information sheets
    (15.5.: deadline for first versions of information sheets; Haifa: concept; Malta: discussion of first versions)
- Joint task: Example analyses for challenging cases
- Interaction with other WGs?
  - With WG 2: Data from MWE templates to use as test sequences for parsers
  - With WG 3: Multilingual perspective on MWE inventories
  - With WG 4: Exchange on classification criteria for tree bank annotation

References

Baldwin, Timothy and Kim, Su Nam (2010). Multiword Expressions. In N. Indurkhya and F. J. Damerau (Eds.), *Handbook of Natural Language Processing* (2 ed.)., pp. 267–292. Boca Raton: CRC Press.

Ellsiepen, Emilia (2005). Distributionsbeschränkungen nicht dekomponierbarer idiomatischer Ausdrücke. Bachelor thesis, University of Tübingen. URL: http://www.coli.uni-saarland.de/∼ellsiepe/emilia_bachelor.pdf.

Sag, Ivan A., Baldwin, Timothy, Copestake, Ann, and Flickinger, Dan (2002). Multiword Expressions: A Pain in the Neck for NLP. In A. F. Gelbukh (Ed.), *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics (CICLING 2002)*, London, pp. 1–15. Springer.