PARSEME/ENeL workshop on MWE e-lexicons

Skopje, 5-6 April 2016

Topics

- European Network of e-Lexicography COST action
- PARSEME/ENeL workshop on MWE e-lexicons
- ELEXIS (European Lexicographic Infrastucture)
  - Call: Integrating Activities for Starting Communities, INFRAIA-02-2017

ENeL COST Action (www.elexicography.eu)

- Institute of the **Polish** Language; Institute of **Czech** Language; Ľ. Štúr Institute, **Slovakia**; Institute of **Croatian** Language and Linguistics; Institute for **Bulgarian** language; **Austrian** Academy of Sciences; Institut für **Deutsche** Sprache, Germany; Elhuyar Foundation, **Basque** Country; Institute for **Dutch** Lexicology; Oxford **English** Dictionary, Oxford University Press; **Fryske** Akademy, Netherlands; Institute for the Languages of **Finland**; Kunnskapsforlaget, **Norway**; Fran Ramovš Institute of the **Slovenian** Language; Árni Magnússon Institute for **Icelandic** Studies; Institute of the **Estonian** Language; Society for **Danish** Language and Literature; **Serbian** Language Institute of SASA; , Institute of **Latvian** Language; …

Working Groups / Objectives

- WG1: Integrated interface to European dictionary content
  - http://www.dictionaryportal.eu/
- WG2: Retro-digitized dictionaries
- WG3: Innovative e-dictionaries
- WG4: Lexicography and lexicology from a pan-European perspective

Innovative e-dictionaries

- The third working group will focus on the development of digitally born dictionaries, focusing on the latest developments in e-lexicography and the interface between **lexicography** and **computational linguistics**.
- Work will be carried out on:
  - the analysis of the possible impact of **automatic acquisition** of lexical data
  - the analysis of the interface between **dictionary** and computational **lexica** (cf. wordnets) and syntactically and semantically **annotated corpora** (cf. FrameNet, SemCor, Senseval)
  - the investigation of the possible use of **dictionary content** for computational linguistic **applications**

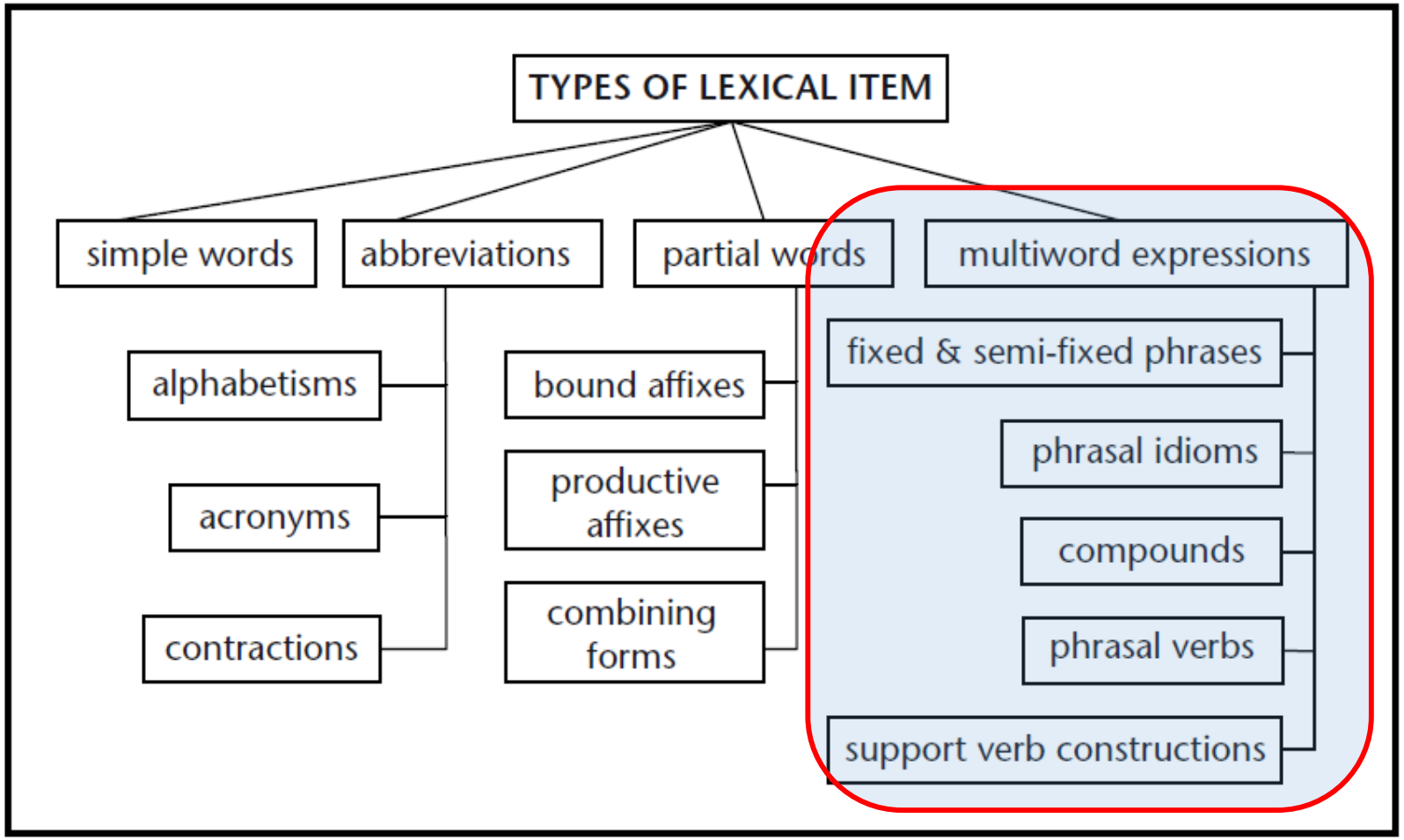
PARSEME/ENeL workshop on MWE e-lexicons

- **Dates**: 5-6 April 2016 (co-located with PARSEME's 6th general meeting on 7-8 April in Struga)
- **Location & Hosting Institution**: University of Skopje, Faculty of Computer Science and Engineering (FCSE), Skopje, FYR of Macedonia (Katerina Zdravkova)
- **Organizers**: Simon Krek, Carole Tiberius, Carla Parra Escartín, Agata Savary
- **Web site**: http://typo.uni-konstanz.de/parseme/index.php/2-general/135-enel-parseme-workshop-on-mwe-lexicons

Tuesday morning

9:00-9:30	**Presentation** **Welcome by workshop leaders and presentation round (max. 1 min. per participant)**
9:30-10:30	**Session 1: MWEs from a lexicographical perspective - a global view** Construction of MWE lexicographical resources. Presentation of frameworks and standards used in Lexicography. Some demos of how lexicographers work Invited speakers: **Lut Colman**, Institute for Dutch Lexicology, the Netherlands **Polona Gantar**, University of Ljubljana, Slovenia
11:00-12:00	**Session 2: MWEs from a Natural Language Processing perspective - a global view** Brief introduction to some NLP tools and applications. How can MWE lexicons improve them? How can NLP tools help in MWE lexicon construction? Invited speaker: **Héctor Martínez Alonso**, University of Paris 7 - INRIA (France)

Practical guide to lexicography (Atkins & Rundell 2008)



Atkins & Rundell 2008: 164

Different MWE classifications and types

Atkins & Rundell (2008)

- fixed and semi-fixed phrases
  - collocations *dock/oak/lettuce leaf*
  - fixed phrases *ham and eggs*
  - similes *drunk as lord*
  - catch phrases *horses for courses*
  - proverbs *too many cooks ...*
  - quotations *to be or not to be*
  - greetings *good morning*
  - phatic phrases *have a nice day*
- (other) phrasal idioms *to have a heart of gold*
- compounds
  - figurative *lame duck*
  - semi-figurative *high school*
  - functional *police dog*
- compound prepositions *in spite of*
- phrasal verbs *get up, hold over, see trough*
- support verb constructions *to take a decision*

Bergenholtz & Gouws (2013)

- collocations
- comparative MWEs
- twin formula *day and night*
- winged words *One small step for man ...*
- routine formula *how do you do*
- proverbs
- MWEs from foreign language *ad hoc*
- expletive constructions *give him an inch and ...*
- MWEs with syntactic function *with regard to*
- idioms *to have eyes in the back of one's head*
- (non-)idiomatic MWEs with a unique component *to and fro*
  - MWEs with an old inflexion
- semi terms *magic eye*
- (non-)idiomatic particle verb
- (non-)idiomatic/non-idiomatic reflexive verb
- noun phrase with semantically void verb

Baldwin & Kim (2010)

- collocations
- idioms
  - verb-noun idiomatic combinations *kick the bucket*
- sentence-like units
- prepositional phrases *in bed, in jail*
  - complex prepositions *in addition to*
- nominal compounds *golf club*
  - complex nominals
- verbal MWEs
  - verb-particle constructions
  - prepositional verbs
  - light-verb constructions

Tuesday afternoon

13:00- 15:00	**Session 3: PARSEME meets ENeL** Plenary interactive demos of MWE resources and tools of the ENeL community Sketch engine, presented by **Miloš Jakubiček** MWE dictionaries and more: practical session by **Lut Colman** and **Polona Gantar**
15:30- 18:00	**Session 4: ENeL meets PARSEME** Plenary interactive demos of MWE-aware NLP tools of the PARSEME community • mwetoolkit, presented by Carlos Ramisch • LeXimir, presented by Cvetana Krstev • Unitex, presented by Matthieu Constant • Demo session of MWE-aware NLP tools

Data sets from participants

- **Idion** (Modern Greek), Stella Markantonatou (ILSP, Athens) - http://www.idion.ilsp.gr
- **Polytropon** (Modern Greek), Voula Giouli (Institute for Language and Speech Processing) - http://goo.gl/SgQlxS at http://athena.clarin.gr/
- **Kollokationenwörterbuch** (Swiss Geman), Tobias Roth (Schweizerisches Idiotikon , Zürich) - http://www.kollokationenwoerterbuch.ch/web/
- **Kamusi** (different languages), Martin Benjamin (Kamusi.org) - https://kamusi.org/
- **Elhuyar MWE resources** (Basque), Antton Gurrutxaga (Elhuyar Foundation, Usurbil)
  https://www.elhuyar.eus/en/site/community/nor-gara-en/fundazioa-en
- **Valency Database of Croatian**, Matea Birtić (Institute for Croatian Language and Linguistics, Zagreb) - http://ihjj.hr/projekt/baza-hrvatskih-glagolskih-valencija/27/
- **Slovene Lexical Database**, Polona Gantar (University of Ljubljana)
  http://eng.slovenscina.eu/spletni-slovar/leksikalna-baza
- **ANW** (Dutch), Lut Colman (INL, Leiden) - http://anw.inl.nl

Wednesday

9:00-10:30	**Session 5: Towards ENeL/PARSEME synergies - hands-on work I** Working in groups with people from both communities • Drafting specifications of an ideal NLP-supported lexicographic framework • Designing a prototype interface with requested functionalities
11:00-12:30	**Session 6: Towards PARSEME/ENeL synergies - hands-on work II** Working in groups with people from both communities • Drafting specifications of an ideal MWE resource that would be useful in different NLP applications • Sample data sets with the desired properties
13:30-15:00	**Final discussion and conclusions** Comparing specifications from sessions I and II and discussing the reasons of divergences • Planning a joint publication

Wish lists

- ENeL (session 5) - A wish list for an ideal NLP-supported lexicographic framework with requested functionalities. This wish list may include:
  - ideal functionalities of MWE tools used for lexicographic purposes
  - prototype interface for MWE tools used for lexicographic purposes
  - description how to include MWE tools in lexicographic workflow
  - other ideas how to use MWE tools for lexicography
- Parseme (session 6) - a wish list for an ideal MWE resource. This wish list should include:
  - Encoded features (word order, morphology, meaning, translation equivalent, reformulation, etc.)
  - Format (input/output) & standard
  - License
  - Target applications (parsing, machine translation, word sense disambiguation, speech recognition, computer assisted language learning, etc.)

Hands-on work

- Session 5 – ideal NLP-supported lexicographic framework
  - **Group 1: identification of known MWEs in corpora**
  - **Group 2: extraction and classification of unknown MWEs**
  - **Group 3: interface for the above tasks within the lexicographic workflow**
- Session 6 – ideal MWE resource
  - **Group 1: MWE lexicons for speech applications**
  - **Group 2: MWE lexicons for monolingual written text application**
  - **Group 3: MWE lexicons for bilingual and multilingual applications**

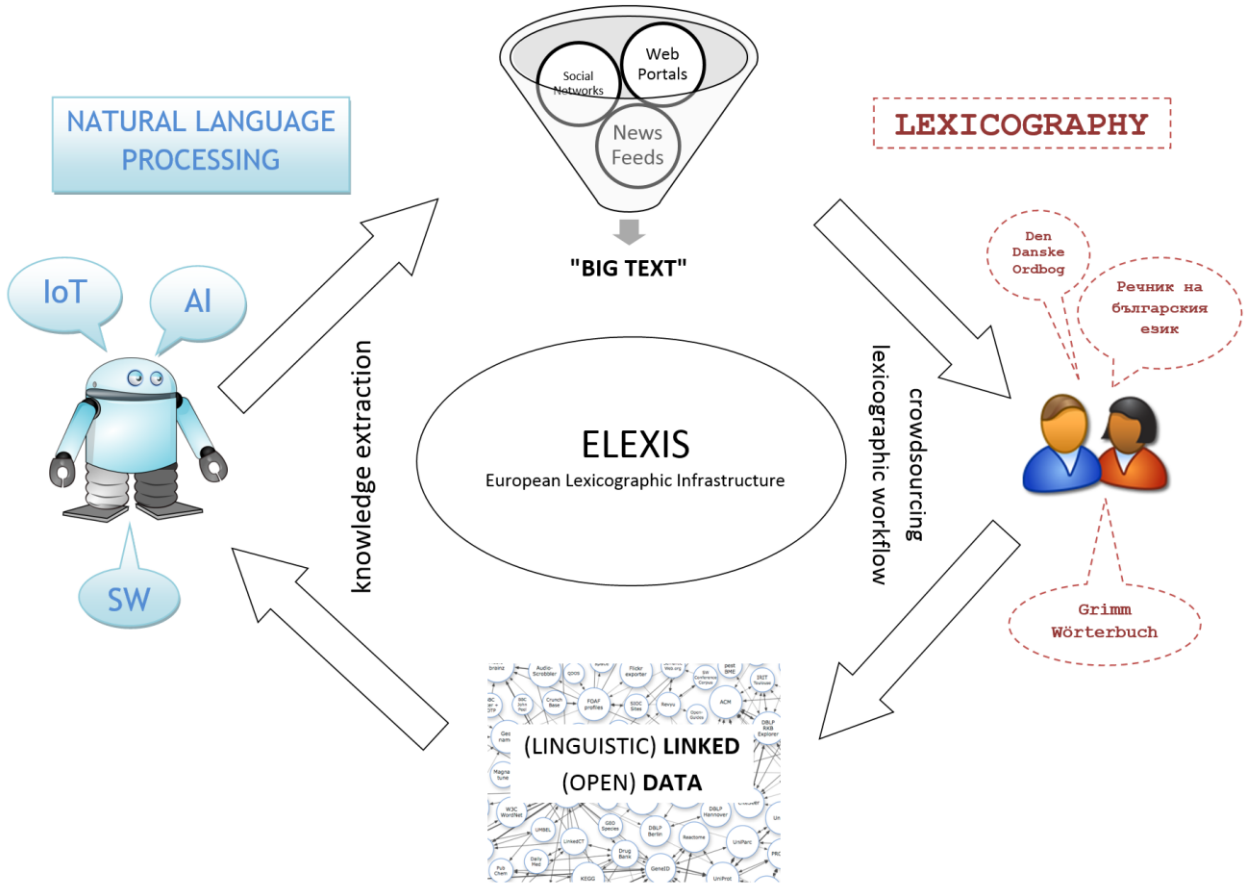
ELEXIS (European Lexicographic Infrastructure)

- Call: INFRAIA-02-2017
- Types of action: RIA Research and Innovation action
- Deadline Model: two-stage
- 1st stage Deadline: 30 March **2016**
- 2nd stage Deadline: 29 March **2017**
- Work Programme Part: European Research Infrastructures (including e-Infrastructures)

Abstract

- The project proposes to integrate, extend and harmonise national and regional efforts in the field of lexicography, both modern and historical, with the goal of creating a sustainable infrastructure which will
- enable efficient **access to high quality lexical data** in the digital age, and
- bridge the gap between more **advanced and lesser-resourced scholarly communities** working on lexicographic resources.
- The need for such an infrastructure has clearly emerged out of the lexicographic community within the European Network of e-Lexicography COST Action which will end in 2017.

Virtuous cycle of e-lexicography



ELEXIS

European Lexicographic Infrastructure