D3.7

Lexical-semantic analytics for NLP: Diachronic distribution of senses - software

Author(s): Federico Martelli (Sapienza), Roberto Navigli (Sapienza), Paola Velardi (Sapienza)

Date: 28/07/2022

H2020-INFRAIA-2016-2017

Grant Agreement No. 731015

ELEXIS - European Lexicographic Infrastructure

Deliverable Number: D3.7

Dissemination Level: Public

Delivery Date: 31/07/2022

Version: 1.0

Author(s): Federico Martelli (Sapienza), Roberto Navigli (Sapienza), Paola Velardi (Sapienza)

2This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Project Acronym:	ELEXIS
Project Full Title:	European Lexicographic Infrastructure
Grant Agreement No.:	731015

Deliverable/Document Information

Project Acronym:	ELEXIS
Project Full Title:	European Lexicographic Infrastructure
Grant Agreement No.:	731015

Document History

Version Date	Changes/Approval	Author(s)/Approved by

4This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Table of Contents

List of Figures

1 Diachronic distribution of senses (software)

The present software deliverable D3.7 is focused on the development of a pipeline which enables the analysis of the evolution of language over the course of centuries and specifically the diachronic distribution of senses. In fact, some senses appeared recently such as the sense of the noun mouse as an electronic device. This phenomenon is crucial when dealing with Word Sense Disambiguation (WSD). Importantly, this work clearly shows that deeper connections between two highly related fields namely Natural Language Processing and lexicography can be mutually beneficial and encourages further collaboration in this direction (Martelli et al. 2021).

The software package which we used for this task is composed of two main components. The first component is a state-of-the-art multilingual WSD system called AMuSE (Orlando et al. 2021). Instead, the second component is a Python script which allows for an analysis of the distribution senses over time. We now detail the usage of our software package.

1.1 Multilingual WSD system

We use the AMuSE multilingual WSD system to disambiguate the corpora which were specifically constructed for this task as described in deliverable D3.7. AMuSE can be used online and offline. To use it please follow the guidelines: http://nlp.uniroma1.it/amuse-wsd/api-documentation. The online API can be queried using this URL: **http://nlp.uniroma1.it/amuse-wsd/api/model**, whereas the offline API can be downloaded and used via a docker image available at: http://nlp.uniroma1.it/resources/ after registering and following the provided instructions. Both online and offline APIs take the following input:

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Method: POST

Request:

Parameter	Type	Description
documents	List<Document>	A list of documents to disambiguate.

1. Document

Attribute	Description
text	**The document to disambiguate.**
lang	**Language of the document.**

We report two usage examples below.

Example (online API):

```
curl -X 'POST' \
  'http://nlp.uniroma1.it/amuse-wsd/api/model' \
  -H 'accept: application/json' \
  -H 'Content-Type: application/json' \
  -d '[
  {"text":"The quick brown fox jumps over the lazy dog.", "lang":"EN"},
  {"text":"I walked along the river bank.", "lang":"EN"}
]'
```


This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Example (offline API):

```
curl -X 'POST' \
  'http://127.0.0.1/api/model' \
  -H 'accept: application/json' \
  -H 'Content-Type: application/json' \
  -d '[
  {"text":"The quick brown fox jumps over the lazy dog.", "lang":"EN"},
]'
```

While the online API covers 10 languages, the offline version includes 40 languages. For further instructions, please visit: http://nlp.uniroma1.it/amuse-wsd/api-documentation

1.2 Software for sense extraction and analysis

The sense-labeled corpora are made available in JSON format as illustrated in Fig. 1.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

```
{
      "wsd_annotations":
                  [
                              {
                                          "bnSynsetId": "bn:00035907n",
                                          "index": 1,
                                          "lemma": "type",
                                          "nltkSynset": "kind.n.01",
                                          "pos": "NOUN",
                                          "text": "types",
                                          "wnSynsetOffset": "5839024n"
                              }
                  ]
}
```


Figure 1 - Data format of the disambiguated corpora

In order to enable and facilitate a diachronic analysis, we created Python scripts which extract a sample of (word, part of speech) pairs which show variations in terms of most frequent sense in the considered time periods. Importantly, each time period might refer to a century or a specific year range indicated by lexicographic partners participating in the ELEXIS project (see D3.7). The Python scripts are made available and documented on the ELEXIS GitHub repository: https://github.com/elexis-eu/diachronic_distribution_of_senses.

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.

Finally, we used this code to perform a thorough diachronic analysis outlined in D3.7.

References

Martelli, F., Navigli, R., Krek, S., Kallas, J., Gantar, P., Koeva, S., Nimb, S., Sandford Pedersen, B., Olsen, S., Langemets, M., Koppel, K., Üksik, T., Dobrovoljc, K., Ureña-Ruiz, R., Sancho-Sánchez, J., Lipp, V., Váradi, T., Győrffy, A., László, S., Quochi, V., Monachini, M., Frontini, F., Tiberius, C., Tempelaars, R., Costa, R., Salgado, A., Čibej, J., & Munda, T. (2021). Designing the ELEXIS Parallel Sense-Annotated Dataset in 10 European Languages. In *Proceedings of eLex 2021.*

Orlando, R., Conia, S., Brignone, F., Cecconi, F., & Navigli, R. (2021, November). AMuSE-WSD: An all-in-one multilingual system for easy Word Sense Disambiguation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 298-307).

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union.