# UFSAC: Unification of Sense Annotated Corpora and Tools

Loïc Vial, Benjamin Lecouteux, Didier Schwab

# UFSAC: Unification of Sense Annotated Corpora and Tools

Loïc Vial  Benjamin Lecouteux  Didier Schwab

{loic.vial,benjamin.lecouteux,didier.schwab}@univ-grenoble-alpes.fr

Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France

## Sense Annotated Corpora

**Sense Annotated Corpora** are **rares** and **costly** resources, and yet they are **essential** for training and/or evaluating **Word Sense Disambiguation** systems.

In English, there are a dozen of manually annotated sense annotated corpora, but their file formats are very different from one another. As a consequence, their usage is often laborious, and some of them are simply never used.

Our work is on the **unification** of these corpora in a **format** that is **easy to use** and **easy to understand**, in order to facilitate the creation of new WSD systems, and the evaluation of existing ones.

Our lexical resource consists of the whole set of sense annotated English corpora in this unified format, along with scripts for building these corpora from their original data, and a complete Java API for manipulating corpora in this format.

The whole resource is available at the following URL: `https://github.com/getalp/UFSAC`

## Statistics of Sense Annotated Corpora in English

| Corpus | Sentences | Words | | Annotated parts of speech | | | |
|---|---|---|---|---|---|---|---|
| | | Total | Annotated | Nouns | Verbs | Adj. | Adv. |
| SemCor [7] | 37176 | 778587 | 229517 | 87581 | 89037 | 33751 | 19148 |
| DSO [11] | 178119 | 5317184 | 176915 | 105925 | 70990 | 0 | 0 |
| WordNet GlossTag [6] | 117659 | 1634691 | 496776 | 232319 | 62211 | 84233 | 19445 |
| MASC [4] | 34217 | 596333 | 114950 | 49263 | 40325 | 25016 | 0 |
| OMSTI [14] | 820557 | 35843024 | 920794 | 476944 | 253644 | 190206 | 0 |
| Ontonotes [3] | 21938 | 435340 | 52263 | 9220 | 43042 | 0 | 0 |
| Senseval 2 [2] | 238 | 5589 | 2301 | 1061 | 541 | 422 | 277 |
| Senseval 3 task 1 [13] | 300 | 5511 | 1957 | 886 | 723 | 336 | 12 |
| SemEval 2007 task 07 [10] | 245 | 5637 | 2261 | 1108 | 591 | 356 | 206 |
| SemEval 2007 task 17 [12] | 120 | 3395 | 455 | 159 | 296 | 0 | 0 |
| SemEval 2013 task 12 [9] | 306 | 8142 | 1644 | 1644 | 0 | 0 | 0 |
| SemEval 2015 task 13 [8] | 138 | 2638 | 1053 | 554 | 251 | 166 | 82 |

## Experiments

We demonstrate the utility of the UFSAC corpora by **extending** a **knowledge-based WSD system** based on the **Lesk** measure.

For every UFSAC corpus and for every word which is sense-annotated in a sentence, we **added** to the **definition** of this sense in the dictionary **every** other **word** present in the sentence.

The resulting definition overlap measure is then evaluated on two tasks of SemEval 2007 and SemEval 2015 evaluation campaigns.

| System | SE 2007 Task 07 | SE 2015 Task 13 |
|---|---|---|
| Lesk [5] | 68.70% | 50.65% |
| Extended Lesk [1] | 78.01% | 61.42% |
| Lesk + UFSAC corpora | 79.83% | 66.43% |
| Most Frequent Sense Baseline | 78.90% | 67.10% |

## UFSAC Format Specifications

The **UFSAC format** is able to contain all the essential information contained in the original formats.

These information are compressed and generalized to the following concepts:

- A **corpus** is the "root" lexical entity which contains a set of documents
- A **document** is a lexical entity which contains a set of paragraphs
- A **paragraph** is a lexical entity which contains a set of sentences
- A **sentence** is a lexical entity which contains a set of words
- A **word** is a "leaf" lexical entity
- Every **lexical entity** may contain a set of **annotations**

The format is based on **XML** and it follows simple conventions:

- **XML nodes** represent **lexical entities**
- **XML attributes** represent **annotations**

**Annotations** include but are not limited to:

- The **surface form** (`surface_form`) of a word
- The **lemma** (`lemma`) of a word
- The **part of speech** (`pos`) of a word
- The **sense** of a word in a specific lexical database, for instance WordNet 3.0 (`wn30_key`)

## UFSAC API

A **Java API** is provided and it allows to easily work with UFSAC corpora.

The **core** package contains the classes **Corpus**, **Document**, **Paragraph**, **Sentence** and **Word** which allow to save/load a corpus to/from a file as well as edit lexical entities and their annotations in memory.

The **streaming** package allow to read, modify and write any corpus file in place without loading it entirely in memory.

## UFSAC Scripts

A set of **useful scripts** is also provided and allow to:

- **Convert a corpus** from its original format to UFSAC
- **Estimate** the **most frequent senses** of all dictionary's lemma on one or many corpora
- **Add POS** and **lemma** annotations to a corpus
- **Evaluate** a **WSD** system by comparing its sense annotations to a gold standard and computing the resulting precision, recall and F1 scores
- **Compute** the score of the **Most Frequence Sense** and **Random baselines** on an evaluation corpus
- **And much more !**

## References

[1] Satanjeev Banerjee and Ted Pedersen. "Extended gloss overlaps as a measure of semantic relatedness". In: *In Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*. 2003, pp. 805–810.

[2] Philip Edmonds and Scott Cotton. "SENSEVAL-2: Overview". In: *The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems*. Toulouse, France: Association for Computational Linguistics, 2001, pp. 1–5.

[3] Eduard Hovy et al. "OntoNotes: The 90% Solution". In: *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. NAACL-Short '06. New York, New York: Association for Computational Linguistics, 2006, pp. 57–60. URL: http://dl.acm.org/citation.cfm?id=1614049.1614064.

[4] Nancy Ide et al. "MASC: the Manually Annotated Sub-Corpus of American English". In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. http://www.lrec-conf.org/proceedings/lrec2008/. Marrakech, Morocco: European Language Resources Association (ELRA), 2008. ISBN: 2-9517408-4-0. URL: http://www.lrec-conf.org/proceedings/lrec2008/pdf/617_paper.pdf.

[5] Michael Lesk. "Automatic sense disambiguation using MRD: how to tell a pine cone from an ice cream cone". In: *Proceedings of SIGDOC '86*. Toronto, Ontario, Canada: ACM, 1986, pp. 24–26. ISBN: 0-89791-224-1.

[6] George A. Miller. "Wordnet: A Lexical Database". In: *ACM* Vol. 38.No. 11 (1995), p. 1–41.

[7] George A. Miller et al. "A semantic concordance". In: *Proceedings of the workshop on Human Language Technology*. HLT '93. Princeton, New Jersey: Association for Computational Linguistics, 1993, pp. 303–308. ISBN: 1-55860-324-7. DOI: 10.3115/1075671.1075742. URL: http://dx.doi.org/10.3115/1075671.1075742.

[8] Andrea Moro and Roberto Navigli. "SemEval-2015 Task 13: Multilingual All-Words Sense Disambiguation and Entity Linking". In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Denver, Colorado: Association for Computational Linguistics, 2015, pp. 288–297. URL: http://www.aclweb.org/anthology/S15-2049.

[9] Roberto Navigli, David Jurgens, and Daniele Vannella. "SemEval-2013 Task 12: Multilingual Word Sense Disambiguation". In: *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. 2013, pp. 222–231. URL: http://www.aclweb.org/anthology/S13-2040.

[10] Roberto Navigli, Kenneth C. Litkowski, and Orin Hargraves. "SemEval-2007 Task 07: Coarse-Grained English All-Words Task". In: *SemEval-2007*. Prague, Czech Republic, 2007, pp. 30–35.

[11] Hwee Tou Ng and Hian Beng Lee. *DSO Corpus of Sense-Tagged English*. 1997.

[12] Sameer S. Pradhan et al. "SemEval-2007 Task 17: English Lexical Sample, SRL and All Words". In: *Proceedings of the 4th International Workshop on Semantic Evaluations*. SemEval '07. Prague, Czech Republic: Association for Computational Linguistics, 2007, pp. 87–92. URL: http://dl.acm.org.gate6.inist.fr/citation.cfm?id=1621474.1621490.

[13] Benjamin Snyder and Martha Palmer. "The English all-words task". In: *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*. 2004. URL: http://www.aclweb.org/anthology/W04-0811.

[14] Kaveh Taghipour and Hwee Tou Ng. "One Million Sense-Tagged Instances for Word Sense Disambiguation and Induction". In: *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*. Beijing, China: Association for Computational Linguistics, 2015, pp. 338–344. URL: http://www.aclweb.org/anthology/K15-1037.