Towards a technology for dictionary intermediated dynamic alignment of multilingual corpora

Dan Cristea\textsuperscript{1,2}, Eveline Wandl-Vogt\textsuperscript{3}, Mihaela Onofrei\textsuperscript{2}, Andrei Scutelnicu\textsuperscript{1,2}

\textsuperscript{1} “Alexandru Ioan Cuza” University of Iași, Faculty of Computer Science, 16 Berthelot St. Iași
\textsuperscript{2} Institute of Computer Science, the Iași branch of the Romanian Academy, 2 Codrescu St. Iași
\textsuperscript{3} Austrian Centre for Digital Humanities (ACDH), Austrian Academy of Sciences Sonnenfelsgasse, 19 – A-1010 Wien

E-mail: dcristea@info.uaic.ro, eveline.wandl-vogt@oeaw.ac.at, mihaela.plamada.onofrei@gmail.com, andreiscutelnicu@gmail.com

Abstract

The paper proposes a methodology for linking large multilingual resources (corpora) with dictionaries and aligning words with a repository of word senses. The proposed method is intended to complement a technology of acquisition of linguistic data in a continuous manner, in order to feed with linguistic material readable dictionaries. Site effects of this basic endeavour are exemplified. The data are for German and Romanian, but the envisioned technology can be applied to any other language.

Keywords: multilingual textual data; aligned linguistic resources; corpora; dictionaries

1. Introduction

In this paper we touch the issue of automatically linking large collections of running language, such as corpora, with dictionaries, in a multilingual perspective. Tremendous advances have been acquired recently in linking linguistic open data (LLOD) (Chiarcos \textit{et al.}, 2013), but a coherent initiative of building a technology able to automatically keep updated a huge collection of multilingual language data is still a task for the future. Moreover, supposing such a technology will be implemented, its benefits would be maximized if the corpora would be aligned among them as well as with existing dictionaries.

2. Motivations and goals

It is widely recognised among lexicographers that dictionaries should have an intimate connection with raw textual data, where word senses should find relevant examples and where new entries and new word senses would have to be detected. The reverse link from corpora to dictionaries is also relevant, because words in contexts should actually be linked to their sense definitions, as those kept in dictionaries. Moreover, seen from a multilingual perspective, linguistic concepts
should have language independent representations, to which sense-identical lexicals belonging to different languages are to be aligned. Specific concepts, particular to only families of languages or language singletons, should also be represented in these repositories of conceptual linguistic data, with obviously fewer or sparse links to different language dictionary entries.

In a perfect world, we can imagine language corpora being accumulated on a daily basis, to keep with language evolution. In these dynamic repositories of linguistic data, the links between words and dictionaries are being permanently updated, to the point that new entries and/or new senses in dictionaries, with corresponding definitions, are being generated when necessary. The technology supporting these evolving multilingual corpora should be able to detect new words and their senses. An universal inventory of linguistic concepts, such is the LOD version of Wiktionary\(^1\) (Declerck et al., 2014), consistently connected onto each specific language dictionary, is perhaps the perfect environment to keep alignments between languages.

### 3. Previous initiatives and resources

In (Cristea, 2010; 2011) a methodology to raise and keep updated an evolving corpus of a language (Romanian in these papers) is proposed. The model describes a curating flow in which textual documents, received from many raw data contributors (editing houses, media channels, universities, etc.), are primarily processed (cleared), then paired with metadata and persistent identification codes and finally sent to a processing chain that does minimally: tokenization, part-of-speech tagging, lemmatization and indexing.

Since the mentioned proposal, a project to build a large Romanian corpus (500 million words) has been initiated by the Romanian Academy: COROLA - the Representative Corpus of Contemporary Romanian Language\(^2\). Moreover, this initiative enjoys also a German-Romanian collaboration within the DRuKoLa project\(^3\) (Cosma et al, 2016). The intention in DRuKoLa is to apply similar principles and accessing technologies to DeReKo, the German Reference Corpus (Deutsches Referenzkorpus), developed at IDS since its inception in 1964 (Kupietz et al., 2010), and COROLA, the Romanian corpus. The querying machinery will mainly be based on the German KORAP platform (Bański et al., 2013). This will allow for cross-linguistic investigations on language-specific grammatical and semantic properties for both German and Romanian.

\(^1\)http://wiktionary.dbpedia.org/

\(^2\) A joint priority project of the Institute of Research in Artificial Intelligence of the Romanian Academy (RACAI), Bucharest, and the Institute of Computer Science of the Iasi branch of the Romanian Academy (ARFI-IIT).

\(^3\) Funded by the Humboldt Foundation and including IDS Mannheim, University of Bucharest, RACAI and ARFI-IIT as partners.
4. Methodology

The delivery of the Romanian corpus, expected to happen at the end of 2017, similarly organised and engined by the same query system as DeReKo, will make these two corpora comparable and will invite to an exciting range of new research. One challenge is to have these two corpora aligned at word senses. This means putting in evidence chains of word occurrences having identical senses, both in-corpora and across-corpora. As such, our proposal promotes a methodology in which more languages will interlink their representative corpora through a common repository of word senses. This endeavour means first that each newly acquired document in the corpus of its specific language should be morphologically tagged and lemmatised, conforming to its specific repository of lemmas. Based on the corpus already acquired, the today technology can then apply vector space models for detecting foreign words, for signalling new words in the newly acquired texts, for recognising senses of words in context and for detecting new senses, with respect to a universal sense inventory, one example being Wiktionary, already mentioned. Out of these, global statistics, tracing the corpora over long periods of time, will be able to signal forgotten (obsolete) words and senses which are no more used.

Once texts are sense tagged with respect to a unique, multilingual, repository of linguistic concepts, a sense detection algorithm could be trained, by taking into consideration signatures of senses in the vicinities of unknown words (del Tredici and Bel, 2015). Similarly, recognition of new linguistically expressed concepts, missing from the sense repositories, can be tried, with a semi-automatic induction of definitions (Labropoulou et al., 2000).

5. Outcomes and discussion

As the first steps in the described direction, we intend to search occurrences of plants, colours and food in the existent COROLA and KORAP indexes, by looking for these categories in the Romanian and the German wordnets (Tufiş et al., 2004; Hamp and Feldweg, 1997) and making use of the rather scarce proliferation of multiple senses in these semantic categories (one word = one, maximum two, senses). Similar entries can then be retrieved from three other very different resources, namely: a) a non-standard, lexicographically curated resource on Bavarian Dialects, the Database of Bavarian Dialects of Austria4, b) a multidisciplinary botanical taxonomy, which is a lexicologically and lexicographically curated resource: the Biodiversity and linguistic diversity portal5 for examples of plants, and c) an aggregated resource, consisting as well of collaboratively curated resources such as Wikipedia and BabelNet6. Out of these first occurrences, contexts could then be learned and a recogniser - trained. As

---

4 https://exploreat.usal.es/
5 https://reconcile.eos.arz.oeaw.ac.at/
6 http://babelnet.org/
such, in a second phase, new colours - plants - food occurrences could be found, e.g. non-standard terms and varieties, based on which the wordnets themselves could be enriched. The found occurrences could thus be chained with similar ones in the pair corpus. Then, we expect that contexts of occurrences belonging to these categories in the two languages could be generalised, on the basis that "semantics" of contexts should be more or less similar.

6. Conclusions

Out of these first experiments, we expect to set aright details of a technology that would allow to go on a much larger scale with the interlingual alignment of corpora intermediated by dictionary entries and their senses. We believe the sketched direction of research, although necessitates a lot of efforts, is very promising. Good results will certainly trigger new approaches towards multidisciplinary, collaborative, cultural lexicography.

7. Acknowledgements

Part of the work described in this paper is supported by the projects COROLA and DRuKoLa.

8. References


