Emmanuel Cartier

DIACHRONIC SEMANTIC EVOLUTION AUTOMATIC TRACKING: A PILOT STUDY IN MODERN AND CONTEMPORARY FRENCH COMBINING DEPENDENCY ANALYSIS AND CONTEXTUAL EMBEDDINGS

Keywords Semantic change; NLP; lexical diffusion; emergence; diffusion and institutionalization; dependency analysis; word embeddings; contextual embeddings

The vocabulary of any given language is continuously evolving: new form-meaning pairs are forged, some signs fall into disuse, and some form-meaning pairs evolve by adding new meanings, losing others or slightly shifting the existing ones. Tracking and describing the dynamism of the vocabulary is one of the main tasks of lexicography. The massive availability of digital or digitalized corpora gives the discipline an unprecedented material for its study, enabling to setup systems able to mine monitor corpora to update the dictionary entries, usages and meanings with specific tools detecting neology and more generally semantic change. However, if it is relatively simple to identify new linguistic signs that appear and those that are no longer used (e.g. among others Kerremans and Prokic 2018; Cartier 2019), it is much more complex to identify semantic neology as it does not manifest at the level of the form itself. Several methods have been proposed by Natural Language Processing (NLP) to try to identify these evolutions of meaning.

Lexical change detection systems have followed advances in NLP methods: after the first systems essentially based on frequency changes (for example Gulordova/Baroni 2011), systems used *word embeddings* (e.g. Kim et al. 2014) and more recently *contextual embeddings* (Hu et al. 2019; Martinc et al. 2019; Giulianelli et al. 2020). These latter systems generally proceed by grouping the contextual vector representations of the different uses into clusters of meaning, then detect changes according to different metrics (Monteirol et al. 2021). Current systems still face many limitations. Mainly, the opacity of neural models does not make it possible to characterize these evolutions, in particular it is difficult, if not impossible, to link the semantic changes to linguistic morphological, syntactic or lexico-syntactic features, or to categorize the types of changes (extension, restriction, metaphor, metonymy, etc.).

To this end, other perspectives have been proposed, based on the hypothesis that meanings are correlated with prototypical co-occurrences or collocations and that an evolution of these elements could denote an evolution of meaning. The most advanced work in this perspective was proposed by (Gries 2012). It is based on the hypothesis that meanings are correlated to prototypical lexical-syntactic patterns (*behavioral profile*). The notion was then extended to that of *dynamic behavioral profile* (Jansegers and Gries 2017) by considering that changes in patterns correlated with semantic evolutions. For example, they show the progressive grammaticalization of one meaning of the verb *sentir*, towards the discourse marker *lo siento* ('I'm sorry') through a visualization by Multidimensional Scaling Maps

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Book of Abstracts of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

(MDS) built from manually annotated linguistic features on contexts of the verb, and a probabilistic model. This method has the disadvantage of a manual annotation, which is time consuming and may bias the results.

In this work, we propose, within the framework of a project aiming at building a reference dataset of semantic evolution in modern and contemporary French (1800–today), from a corpus of journalistic texts from the French National Library website (1850–1940) and a contemporary newspaper corpus from the web (2014–2021) to study a sample of polysemous words (nouns and verbs). To trace their evolution, we explore a combination of the above approaches: on the one hand, *contextual embeddings*, which allow to get a very fine representation of the context and to cluster the vectorized representations of individual occurrences, thus discovering clusters of meaning and their evolution through time; as pretrained model, we use CamemBERT (Martin et al. 2019), a state-of-the-art model for French; on the other hand, a dependency analysis by means of a state-of-the-art parser (Straka 2018), allowing not only to annotate each token with morphological but also syntactic features. The latter approach, inspired by (Jenserges and Gries 2017) does not use any manual annotation, and focus on valid lexico-syntactic patterns for the two categories: for nouns, modifier (N ADJ, N prep N), and core argumental constructions (N subject or object of a verb). For verbs, core argumental constructions (subject group, direct and indirect object, subordinate clause).

In this contribution, we will present the different phases of the project (corpora building, choice of lexemes, pre-processing and exploration web platform) and the first lessons learned.

References

Giulianelli, M./Tredici, M.D./Fernández, R. (2020): Analysing lexical semantic change with contextualised word representations. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, July 5–10, 2020. Stroudsburg, PA, pp. 3960–3973.

Gries, St. Th. (2012): Behavioral profiles: a fine-grained and quantitative approach in corpus-based lexical semantics. In: Jarema, G./Libben, G./Westbury, Ch. (eds.): Methodological and analytic frontiers in lexical research. Amsterdam, pp. 57–80.

Gulordava, K./Baroni, M. (2011): A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In: Proceedings of the GEMS 2011 Workshop on Geometrical Models of Natural Language Semantics. Edinburgh, pp. 67–71.

Hu, R./Li, S./Liang, S. (2019): Diachronic sense modeling with deep contextualized word embeddings: an ecological view. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, pp. 3899–3908.

Iavarone, B./Brunato, D./Dell'Orletta, F. (2021): Sentence complexity in context. In: Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, Online Event, June 10, 2021, pp. 186–199.

Jansegers, M./Gries, St. Th. (2017): Towards a dynamic behavioral profile: a diachronic study of polysemous sentir in Spanish. In: Corpus Linguistics and Linguistic Theory 16 (1), pp. 145-187.

Kim, Y./Chiu, Y.-I./Hanaki, K./Hegde, D./Petrov, S. (2014): Temporal analysis of language through neural language models. In: Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science. Baltimore, pp. 61–65.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Book of Abstracts of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Kosem, I./Koppel, K./Kuhn, T. Z./Michelfeit, J./Tiberius, C. (2019): Identification and automatic extraction of good dictionary examples: the case(s) of GDEX. In: International Journal of Lexicography 32, pp. 119–137.

Martin, L./Muller, B./Ortiz Suarez, P./Dupont, Y./Romary, L./Villemonte de la Clergerie, E./Seddah, D./Sagot, B. (2020): CamemBERT: a tasty French language model. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, July 5–10, 2020. Stroudsburg, PA, pp. 7203–7219.

Martinc, M./Novak, P. K./Pollak, S. (2020): Leveraging contextual embeddings for detecting diachronic semantic shift. In: Proceedings of the 12th Conference on Language Resources and Evaluation (LREC), Marseille, 11–16 May 2020, pp. 4811–4819.

Montariol, S./Martinc, M./Pivovarova, L. (2021): Scalable and interpretable semantic change detection. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, June 6–11, 2021, pp. 4642–4652.

Raganato, A./Pasini, T.,/Camacho-Collados, J./Pilehvar, M. T. (2020): XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, November 16–20, 2020, pp. 7193–7206.

Segonne, V./Candito, M./Crabbé, B. (2019): Using Wiktionary as a resource for WSD: the case of French verbs. In: Proceedings of the 13th International Conference on Computational Semantics, Gothenburg, Sweden, pp. 259–270.

Straka, M. (2018): UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task. In: Proceedings of CoNLL 2018: The SIGNLL Conference on Computational Natural Language Learning, Brussels, Belgium, pp. 197–207.

Contact information

Emmanuel Cartier
LIPN – RCLN UMR 7030 CNRS, University Sorbonne Paris Nord (Paris, France)
emmanuel.cartier@univ-paris13.fr

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Book of Abstracts of the XX EURALEX International Congress. Mannheim: IDS-Verlag.