

# Introduction to the special section on the interaction between formal and computational linguistics

*Timothée Bernard*<sup>1</sup> and *Grégoire Winterstein*<sup>2</sup>

<sup>1</sup> Université Paris Cité

<sup>2</sup> Université du Québec à Montréal (UQAM)

## INTRODUCTION

1

While computational linguistics is historically rooted in formal linguistics, it might seem that the distance between these two fields has only grown larger as each field evolved. Still, whether this impression is correct or not, not all links have been cut, and new ones have appeared. Indeed, while we are currently witnessing a growing interest within formal linguistics in both explaining the remarkable successes of neural-based language models and uncovering their limitations, one should not forget the contribution to theoretical linguistics provided, for example, by the computational implementation of grammatical formalisms. And while neural-based methods have recently received the lion's share of the public attention, interpretable models based on symbolic methods are still relevant and widely used in the natural language processing industry.

The links that exist between formal and computational linguistics have been the subject of discussion for a long time. At the 2009 European Meeting of the Association for Computational Linguistics, a workshop entitled “Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?” was organised. This workshop led to the publication a couple of years later of the sixth volume of *Linguistic Issues in Language Technology* (Baldwin and Kordoni 2011). At the centre of this publication were discussions about

how and why formal linguistics and computational linguistics went down different paths, about the benefits and drawbacks of specialisation and about how our scientific communities could improve the situation. On this occasion, Church (2011) predicted that computational approaches to language would come back to symbolic approaches, observing that most of the “low hanging fruits” of statistical methods had already been picked. In a similar vein, Kay (2011) argued that natural language processing (NLP) – distinguished from computational linguistics in that the former was held to show little interest in language, and to be oriented towards pure performance matters – would disappear.

However, far from dwindling, statistical methods garnered renewed interest due to impressive advances in machine learning and, in particular, the progress made in the development of word embeddings generated as a product of the optimisation of neural-based language models (Mikolov *et al.* 2013b,a; Bengio *et al.* 2001). This stream of research eventually led to the apparition of the Transformer architecture (Vaswani *et al.* 2017), with famous implementations such as BERT (Devlin *et al.* 2019) and GPT-3 (Brown *et al.* 2020) which offer linguistic representations that are routinely used for a wide array of NLP applications, from classification to language generation. Though remarkably effective, with benchmark performances regularly smashed by newer and bigger models, the representations offered by these systems are largely shunned by the linguistic community, who often sees them as irrelevant to our understanding of language (see *infra*).

As already mentioned, it would, however, be a little hasty to declare a divorce between linguistic and computational methods. First, using computational methods to validate theoretical models remains common practice in many circles (e.g. among the LFG, HPSG or categorial grammar communities) and the use of such implementations can also be used to investigate and test typological hypotheses about language universals (such as with the LinGo Grammar Matrix; Bender *et al.* 2002). The use of symbolic methods also remains common in the industry, especially for applications for which humanely interpretable models are necessary (for various reasons including ethical ones; see Lipton 2018; Miller 2019 and references therein). Second, the properties of stochastic language models have also come under increasing

scrutiny. On the one hand, there is a lively debate about the ability of these models to properly represent natural language meaning (Bender and Koller 2020), and about how representative they are of the linguistic practices of the members of a linguistic community (Bender *et al.* 2021). On the other hand, there are efforts to explain the sheer effectiveness of these language models, in a way that goes beyond the mere mention of the distributional hypothesis (Gastaldi 2020), and to investigate how such models are sensitive (or not) to complex linguistic phenomena such as presupposition projection (Jiang and de Marneffe 2019) or syntactic generalisations (Hu *et al.* 2020) (see also the domain of “BERTology”, which seeks to study the properties of the representations manipulated by BERT-like models, though not necessarily from a linguistic angle; Ettinger 2020; Rogers *et al.* 2020).

## OVERVIEW OF THE SPECIAL SECTION

2

Inspired by these tensions and connections, we organised a one-day online event on the interactions between formal and computational linguistics which took place in June 2021.<sup>1</sup> The guiding thread for the talks at that event was, roughly, to focus on and discuss recent advances in computational linguistics (be they symbolic or not), their relationship with linguistic data, and what such systems can do for language and linguistics itself. These questions were tackled from different angles: practical, theoretical and philosophical. The present special section takes its roots in that event, as we offered the presenters a chance to elaborate on the themes developed in the workshop in the form of long papers.

Both articles in this special section illustrate the theme of the seminar in two complementary ways, both in their use of computational methods to address theoretical issues of formal models of language,

---

<sup>1</sup>See <https://gdr-lift.loria.fr/news/ilfc-en/>. The event then turned into a monthly online seminar <https://gdr-lift.loria.fr/monthly-online-ilfc-seminar/>.

and in the way they use linguistically inspired symbolic methods to achieve their goals.

In their paper, Olga Zamaraeva and her co-authors retrace the evolution of the “Grammar Matrix”, a meta-grammar engineering framework that relies on the HPSG (Pollard and Sag 1994) and MRS (Copestake *et al.* 2005) frameworks. The Grammar Matrix is a tool that automates the implementation of the grammar for a given language. To do so, the user provides information about the properties of the language they wish to implement a grammar of, along with a sample lexicon. On the basis of those properties and known analyses of the related phenomena in HPSG and MRS, the matrix is able to produce an implemented grammar that can be used, among other things, to test the coverage of the grammar on a set of sentences. Beyond that, the authors also highlight how the Grammar Matrix can be used to investigate cross-linguistic variation, and formulate and test general hypotheses about the structure of language. On the basis of a test set of sentences in 60 different languages from 40 distinct families, a regression testing system is used to check how modification in the analyses of phenomena affect the overall architecture of the system. The Grammar Matrix is thus a prime example of how computational methods can directly influence linguistic analysis, both as a tool to test such analyses, and as a way to get better insight about language using an approach that is both theoretically and empirically grounded.

Haruta *et al.* present the theoretical foundations and the practical implementation of an automatic Natural Language Inference (NLI) solver for English, i.e. a system that, given two input texts, aims at detecting whether the first entails, contradicts or is neutral toward, the second. One characteristic of their solver is that it is *symbolic*; while a recent popular approach in NLI (as in other NLP tasks) consists in training a classifier using only vector representations obtained via a language model (see *supra*), the system they describe relies on logical representations of the input texts produced by a parser and fed to a theorem prover. The various parsers they use are based on the Combinatory Categorical Grammar formalism (CCG; Steedman and Baldridge 2011); after a little bit of post-processing of the output trees, the logical representations are standardly derived from the syntactic analyses in a compositional fashion.

Obviously, the success of the enterprise crucially depends, among other things, on the expressive power of the logical language used. Haruta et al. have chosen to express the semantics of sentences in a version of First Order Logic (FOL) that incorporates events and integers/degrees, allowing them to translate a wide range of constructions involving adjectives, comparatives, generalised quantifiers and numerals. Key to many of their analyses is the notion of degree. For example, they analyse *Tom is taller than Mary* following the A-not-A analysis (see Schwarzschild 2008 and references therein) as meaning that there is some degree such that Tom has, but Mary has not, this degree of tallness. Haruta et al. evaluate their system on a large number of NLI datasets, including a novel one they have designed to cover the phenomena that they have been particularly interested in, usually absent from existing datasets such as FraCas (Cooper *et al.* 1996). Results show that, in general, non-symbolic models perform significantly worse than state-of-the-art symbolic models, and that, in particular, the system presented here is particularly effective. This very interesting paper thus contributes to showing that formal syntax and semantics are still relevant to natural language processing and that, in some domains, symbolic reasoning is still one step ahead of the purely neuronal alternatives that have progressively taken the spotlight in the last decade.

## ACKNOWLEDGMENTS

The workshop and seminar out of which this special section grew was logistically supported by the “Groupe de Recherche en Linguistique Informatique, Formelle et de Terrain” (GdR LIFT, <https://gdr-lift.loria.fr/>), which is a structure funded by the French national scientific research agency in an effort to foster discussion among three communities related to language: the communities of field linguistics, formal linguistics and computational linguistics.

## REFERENCES

- Timothy BALDWIN and Valia KORDONI (2011), The Interaction between Linguistics and Computational Linguistics, *Linguistic Issues in Language Technology*, 6, doi:10.33011/lilt.v6i.1233, <https://journals.colorado.edu/index.php/lilt/article/view/1233>.
- Emily M. BENDER, Dan FLICKINGER, and Stephan OEPEN (2002), The Grammar Matrix: An Open-Source Starter-Kit for the Rapid Development of Cross-Linguistically Consistent Broad-Coverage Precision Grammars, in John CARROLL, Nelleke OOSTDIJK, and Richard SUTCLIFFE, editors, *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, pp. 8–14, Taipei, Taiwan.
- Emily M. BENDER, Timnit GEBRU, Angelina McMILLAN-MAJOR, and Shmargaret SHMITCHELL (2021), On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?, in *Proceedings of FAccT 2021*, pp. 610–623.
- Emily M. BENDER and Alexander KOLLER (2020), Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5185–5198, Association for Computational Linguistics, Online, <https://www.aclweb.org/anthology/2020.acl-main.463>.
- Yoshua BENGIO, Réjean DUCHARME, and Pascal VINCENT (2001), A Neural Probabilistic Language Model, in T. K. LEEN, T. G. DIETTERICH, and V. TRESP, editors, *Advances in Neural Information Processing Systems 13*, pp. 932–938, MIT Press, <http://papers.nips.cc/paper/1839-a-neural-probabilistic-language-model.pdf>.
- Tom BROWN, Benjamin MANN, Nick RYDER, Melanie SUBBIAH, Jared D. KAPLAN, Prafulla DHARIWAL, Arvind NEELAKANTAN, Pranav SHYAM, Girish SASTRY, Amanda ASKELL, Sandhini AGARWAL, Ariel HERBERT-VOSS, Gretchen KRUEGER, Tom HENIGHAN, Rewon CHILD, Aditya RAMESH, Daniel ZIEGLER, Jeffrey WU, Clemens WINTER, Chris HESSE, Mark CHEN, Eric SIGLER, Mateusz LITWIN, Scott GRAY, Benjamin CHESSE, Jack CLARK, Christopher BERNER, Sam MCCANDLISH, Alec RADFORD, Ilya SUTSKEVER, and Dario AMODEI (2020), Language Models are Few-Shot Learners, in H. LAROCHELLE, M. RANZATO, R. HADSELL, M.F. BALCAN, and H. LIN, editors, *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, Curran Associates, Inc., <https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf>.
- Kenneth CHURCH (2011), A Pendulum Swung too Far, *Linguistic Issues in Language Technology*, 6, <https://journals.colorado.edu/index.php/lilt/article/view/1245>.

*Introduction to the special section*

Robin COOPER, Dick CROUCH, Jan VAN EIJCK, Chris FOX, Josef VAN GENABITH, Jan JASPARS, Hans KAMP, David MILWARD, Manfred PINKAL, Massimo POESIO, and Steve PULMAN (1996), Using the Framework, Technical Report Deliverable D16, <https://gu-clasp.github.io/multifracas/D16.pdf>.

Ann COPESTAKE, Dan FLICKINGER, Ivan SAG, and Carl POLLARD (2005), Minimal Recursion Semantics: An Introduction, *Research in Language and Computation*, 3(2–3):281–332.

Jacob DEVLIN, Ming-Wei CHANG, Kenton LEE, and Kristina TOUTANOVA (2019), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Association for Computational Linguistics, Minneapolis, Minnesota, doi:10.18653/v1/N19-1423, <https://www.aclweb.org/anthology/N19-1423>, 04472.

Allison ETTINGER (2020), What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models, *Transactions of the Association for Computational Linguistics*, 8:34–48, doi:doi.org/10.1162/tacla00298.

Juan Luis GASTALDI (2020), Why Can Computers Understand Natural Language?, *Philosophy & Technology*, doi:10.1007/s13347-020-00393-9.

Jennifer HU, Jon GAUTHIER, Peng QIAN, Ethan WILCOX, and Roger P. LEVY (2020), A Systematic Assessment of Syntactic Generalization in Neural Language Models, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1725–1744.

Nanjiang JIANG and Marie-Catherine DE MARNEFFE (2019), Evaluating BERT for Natural Language Inference: A Case Study on the CommitmentBank, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019)*, pp. 6086–6091, doi:10.18653/v1/D19-1630, <https://aclanthology.org/D19-1630>.

Martin KAY (2011), Zipf's Law and *L'Arbitraire du Signe*, *Linguistic Issues in Language Technology*, 6, <https://journals.colorado.edu/index.php/lilt/article/view/1251>.

Zachary C. LIPTON (2018), The Mythos of Model Interpretability, *ACM Queue*, 16(3):1–27, doi:10.1145/3236386.3241340, <http://doi.acm.org/10.1145/3236386.3241340>.

Tomas MIKOLOV, Kai CHEN, Greg CORRADO, and Jeffrey DEAN (2013a), Efficient Estimation of Word Representations in Vector Space, in Yoshua BENGIO and Yann LECUN, editors, *1st International Conference on Learning*

Representations, *ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, <http://arxiv.org/abs/1301.3781>.

Tomas MIKOLOV, Ilya SUTSKEVER, Kai CHEN, Greg S. CORRADO, and Jeff DEAN (2013b), Distributed Representations of Words and Phrases and their Compositionality, in C. J. C. BURGESS, L. BOTTOU, M. WELLING, Z. GHAHRAMANI, and K. Q. WEINBERGER, editors, *Advances in Neural Information Processing Systems 26*, pp. 3111–3119, Curran Associates, Inc., <http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf>.

Tim MILLER (2019), Explanation in Artificial Intelligence: Insights from the Social Sciences, *Artificial Intelligence*, 267:1–38, doi:10.1016/j.artint.2018.07.007, <http://www.sciencedirect.com/science/article/pii/S0004370218305988>.

Carl J. POLLARD and Ivan A. SAG (1994), *Head-Driven Phrase Structure Grammar*, University of Chicago Press, Chicago.

Anna ROGERS, Olga KOVALEVA, and Anna RUMSHISKY (2020), A Primer in BERTology: What We Know About How BERT Works, *Transactions of the Association for Computational Linguistics*, 8:842–866, doi:10.1162/tacl\_a\_00349, <https://aclanthology.org/2020.tacl-1.54>.

Roger SCHWARZSCHILD (2008), The Semantics of Comparatives and Other Degree Constructions, *Language and Linguistics Compass*, 2(2):308–331, doi:10.1111/j.1749-818X.2007.00049.x, <https://onlinelibrary.wiley.com/doi/10.1111/j.1749-818X.2007.00049.x>.

Mark STEEDMAN and Jason BALDRIDGE (2011), Combinatory Categorical Grammar, in Robert D. BORSLEY and Kersti BÖRJARS, editors, *Non-Transformational Syntax*, pp. 181–224, Wiley-Blackwell, ISBN 978-1-4443-9503-7, doi:10.1002/9781444395037.ch5, <http://onlinelibrary.wiley.com/doi/10.1002/9781444395037.ch5/summary>.

Ashish VASWANI, Noam SHAZEER, Niki PARMAR, Jakob USZKOREIT, Llion JONES, Aidan N. GOMEZ, Łukasz KAISER, and Illia POLOSUKHIN (2017), Attention is All you Need, in I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN, and R. GARNETT, editors, *Advances in Neural Information Processing Systems 30*, pp. 5998–6008, Curran Associates, Inc., <http://papers.nips.cc/paper/7181-attention-is-all-you-need>.



*Introduction to the special section*

*Timothée Bernard*

© 0000-0003-4172-6986

timothee.bernard@u-paris.fr

Laboratoire de linguistique formelle

(LLF)

Université de Paris

8 place Paul Ricœur

75205 Paris Cedex 13, France

*Grégoire Winterstein*

© 0000-0002-8951-2138

winterstein.gregoire@uqam.ca


Université du Québec à Montréal

(UQAM)

Timothée Bernard and Grégoire Winterstein (2022), *Introduction to the special section on the interaction between formal and computational linguistics*, *Journal of Language Modelling*, 10(1):39–47

doi <https://dx.doi.org/10.15398/jlm.v10i1.325>

This work is licensed under the *Creative Commons Attribution 4.0 Public License*.

cc  <http://creativecommons.org/licenses/by/4.0/>