# Bringing *NL* Back with *P*: Defending Linguistic Methods in NLP for Future AI Applications

Fabio Meroni[1,2][a]
[1]*X23 – Science in Society, Treviglio, Italy*
[2]*Università degli Studi di Bergamo, Bergamo, Italy*

Keywords: Natural Language Processing, Computational Linguistics, Artificial Intelligence, Natural Language Formalisation, Lexicon-Grammar, NooJ.

Abstract: In the development of Artificial Intelligence (AI) tools for Natural Language Processing (NLP) applications, this position paper promotes the (re)introduction of rule-based, linguistically informed methodologies, with particular attention to addressing the challenges posed by low-resource languages and research ethics when it comes to the enhancement of machine intelligence by means of linguistic intelligence. NLP, as a rapidly evolving subfield of AI, has seen a proliferation of contributions in recent years. However, the predominant reliance on statistically driven approaches has reduced NLP to a pursuit of superficial aesthetic results, neglecting the foundational linguistic structures that underpin natural language processing. Consequently, the marginalization of linguists within the field has stalled progress toward a deeper understanding of Natural Language Formalization (NLF). Without targeted intervention, these issues threaten to persist, undermining the potential of NLP to achieve its full intellectual and practical promise. This paper argues for a renewed integration of the science of natural language (NL) into its processing (P) within an interdisciplinary framework that emphasizes collaboration between computational linguists and AI researchers, and presents a methodological proposition of a possible way to include linguistic resources in a richly informed AI application using NooJ.

## 1 INTRODUCTION

This position paper advocates for the inclusion of rule-based linguistic tools in the development of new Natural Language Processing (NLP) applications. This approach is not forcedly intended to replace AI-driven tools, but rather to complement them by integrating an additional layer of functionality that accounts for the description of actual linguistic resources, for the sake of grammaticality and scientific rigour. After offering an overview of the state-of-the-art in theoretical and methodological approaches to NLP, this contribution presents how computational tools rooted in "linguistic methods" (Silberztein, 2024) bring valuable insights and grant a critical foundation for the development of linguistic intelligence, a key component in achieving the broader objective of "true" artificial intelligence. By combining these methodologies, the following paragraphs show how the industry can envisage building more robust and versatile NLP applications that leverage the strengths of linguistic theory when applied to computational formalisation, in order to achieve higher flexibility in a number of language processing tasks including text generation, spell-checking, automatic Part Of Speech (POS) tagging and machine translation. This proposal insists on how NooJ, as a software and a methodological staple rooted in lexicon-grammar (Gross, 1994), can provide an alternative perspective by emphasising language processing in a representational way rather than a purely statistical one, supporting the linguistic approach to NLP first and foremost when approaching the study of low-resource languages.

## 2 TROUBLED TIMES OF NLP

Stating that NLP primarily aims to formalise the complexities of natural language may appear so self-evident that one might choose to omit it altogether when speaking about this domain of studies. One could

[a] https://orcid.org/0009-0005-5853-6005

Meroni, F.
Bringing NL Back with P: Defending Linguistic Methods in NLP for Future AI Applications.
DOI: 10.5220/0013259100003890
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 3*, pages 1062-1068
ISBN: 978-989-758-737-5; ISSN: 2184-433X
Proceedings Copyright © 2025 by SCITEPRESS – Science and Technology Publications, Lda.

argue that acknowledging this foundational goal is crucial, as it highlights the inherent challenges involved in research accounts made in this field: by recognising the importance of Natural Language Formalisation (NLF), researchers and practitioners should underscore the interdisciplinary nature of NLP, which integrates elements from linguistics and computer science to create models that approximate human understanding, communication, and reasoning.

However, recent accounts in the history of this discipline have been talking about a sort of paradigm shift: after a first phase of absolute reliance on linguistic modelling, a second (ongoing) wave sees AI taking the driving seat, marking a “renaissance” of NLP (see at least Fanni et al., 2023 and Jiang et al., 2023, p. 2). As a consequence, NLP methods that find their very basis in the linguistically-accurate description of phenomena occurring in natural language instantiations and productions are now widely judged to be outdated if not even obsolete. According to the retelling of Anitha S. Pillai and Roberto Tedesco (2024):

> [...] especially after the advent of very powerful conversational agents able to simulate a human being and interact with the user in a very convincing way, AI and the historical field of Natural Language Processing almost become synonymous (think of the abilities of GPT-3-derived models; for example, ChatGPT1). (Pillai & Tedesco, 2024, p. vii).

The same two authors proceed with a very bold statement when it comes to their diachronic overview of research in this field:

> Historically, NLP approaches took inspiration from two very different research fields: linguistics and computer science; in particular, linguistics was adopted to provide the theoretical basis on which to develop algorithms trying to transfer the insight of the theory to practical tasks. Unfortunately, this process proved to be quite difficult, as theories were typically too abstract to be implemented as an effective algorithm. On the other hand, computer science provided plenty of approaches, from AI and Formal Languages fields. (Pillai & Tedesco, 2024, p. x).

The “difficulty” encountered in converting linguistic theory into algorithmic form may not only be a consequence of inherent theoretical complexity, but also of a limited willingness to work with linguists to translate insights into implementable formats. Thus, while there is undeniable value in linguistics as a theoretical cornerstone, a certain reticence to engage in cross-disciplinary adaptation has limited the field's influence.

All in all, the argument presented by Pillai and Tedesco (2023) reflects a subtly dismissive view of the relationship between linguistics and NLP, attributing limited success in early times to the supposed impracticality of linguistic theory, yet failing to account for the potential value of linguistics as the very basis of scientifically relevant results in the field if approached with rigour and adaptability. Even though NLP continues to be defined as a discipline that combines linguistics with AI and computer science, we are now in the conditions to believe NLP is more and more about the latter two than the former, apparently foundational one discipline that strives for scientificity and explainability. In short, linguists have been excluded from the discussion, while the remaining experts at the table are celebrating their “advancements”[1].

This paper hereby highlights the need for greater scholarly agility in an interdisciplinary call to action, with the objective of rendering linguistic theories actionable within computational frameworks. In today’s technological landscape, computational linguistics (CL) must not simply provide “inspiration”, but should actively integrate into AI, offering a scaffold for interpretability, structure, and meaningful context to NLP models. Embracing this shift requires software engineers to re-engage with the field together with their colleagues in language-related studies, to refine theories in ways that make them more computationally relevant without losing the depth of insight that linguistics uniquely offers as is very common nowadays.

This leads me to consider as such the current status of NLP: more and more P, less and less NL.

## 3 THE IMPORTANCE OF DEALING WITH NATURAL LANGUAGE

Time for some recent history, this time in a more positive light. Yogatama et al. (2019), long before the popularisation of AI-powered chatbots (see the now unsurpassably popular ChatGPT by OpenAI,

[1] See for example how in 2024 the scientific journal *Mathematics* released a special issue titled “Current Trends in Natural Language Processing (NLP) and Human Language Technology (HLT)”. Unsurprisingly, no contribution in it has any reference to linguistic theory.

launched in 2022), tempered the enthusiasm for the apparent progress that NLP seemed to be making with the advent of LLMs, emphasising the need to evaluate model performance based on demonstrated linguistic intelligence rather than solely on the plausibility of their output. In fact, overly simplifying in a sense but leaving nothing out of the frame, cutting-edge examples of generative AI produce sentences that resemble meaningful instantiations of natural language, yet at their base they don't follow any rules. In a nutshell, the model is simply reassembling patterns is seen before, based on statistical likelihoods.

Emily Bender and Alexander Koller (2020) added that, since human-analogous processing of natural language is a big goal of research in artificial intelligence, we should reconsider the futuristic narrative that sees LLMs as their protagonists and focus more on questions of machine intelligence regarding language use. The same Emily Bender has been at the centre of the scene since she first-authored the illuminating paper which introduced the concept of "stochastic parrot" (Bender et al., 2021) as a neologism to designate the ignorance that AI shows with respect to any semantic implication of the words they draw from LLMs and subsequently employ; it is not a case that Bender is a computational linguist, not a software engineer, engaged in an ongoing project focused on the production of linguistic formalisms through unrestricted grammars[2]. Also Jiang et al. (2023, p. 43 and sq.) have insisted on the need for NLP AI-driven tools to rely on some sort of rule-based knowledge, including in this proposition of theirs the hunger for a good deal of linguistic information.

## 4 THE STATUS OF NL AND NLP

In his publications, Max Silberztein opposes what he calls "linguistic methods" to "statistical" or – with a peculiar choice of terminology that could clash with shared conventions [3] – "empirical methods" (Silberztein, 2016; 2024), and always advocated for the first. When defending the linguistic approach, he outlined ten reasons why merely stochastic methods should be disregarded (Silberztein, 2016, pp. 19-27), of which the tenth contains the quintessential point of my argumentation: one can't derive from the application of statistical methods, even when they produce high-quality if not spectacular results in terms of NLP tasks, any linguistic information of scientific relevance, but mere mathematical functions. His preface to *Linguistic Resources for Natural Language Processing* (Silberztein, 2024) is a brutally concise manifesto of such ideal.

This proves to be particularly relevant for the study of low-resource languages. In the indigenous-led position paper resulting from the initiative of Lewis et al. (2020, pp. 35-36, p. 65, pp. 93-100), language is a key element around which AI ethics, representation, and inclusivity revolve. The authors propose frameworks for Indigenous data governance and culturally aware AI systems that prioritise indigenous ownership and stewardship of digital linguistic resources over the massive application of statistical computing on transcribed corpora in endangered languages carrying traditional knowledge.

Linguists working in these contexts may face some challenges when they try to introduce digital methodologies in their language documentation tasks, among which the most necessary is interlinearisation, a process of aligning a text with its translation line by line. So far, in the efforts intended to produce NLP tools for linguists working with low-resource languages, cutting-edge research in the field of annotation and interlinearisation of texts has been conducted using machine learning techniques applied to training datasets. Zhao et al. (2020) proposed a model based on Recurrent Neural Networks (RNN) using source transcription and its translation for the automatic gloss generation task, trying to overcome the need for word alignment that is imperative when using Conditional Random Fields (CRF), which had been applied shortly before by Angelina McMillan-Major (2020) for the same purpose. Diego Barriga Martínez and colleagues (2021) have experimented with both RNN and CRF, and described how factors like systematicity and frequency of morphological rules in the training corpus are directly proportional to precision scoring. It is relevant to cite how Moeller et al. (2020) focused on goals of morphological accuracy, and showed a methodology based on transformers that nevertheless includes, in an

[2] See https://linguistics.washington.edu/research/projects-and-grants/lingo-grammar-matrix.

[3] In linguistic theory, "empirical linguistics" can be said to be the study of language based on direct observation and data collection from real-world language use (see at least Sampson, 2002, and Schütze, 2016 [1996]). Many linguists, especially field linguists, would say they take pride in being empirical, as their empirical studies would involve qualitative methods, such as discourse analysis or ethnographic observations; in his terminological choice, however, Silberztein only refers to the quantitative side of empirical analysis, like statistical modelling of language-use samples recollected in training corpora. On the other hand, he calls "linguistic" all rule-based methods.

intermediate passage, human intervention: after partial inflectional paradigms were automatically extracted from the interlinear glossed texts, the linguist Andrew Brumleve was asked to "clean" the resulting data and regroup words under common features and inflectional paradigms, implementing corrections where necessary (Moeller et al., 2020, p. 5256); such an operation of fine-tuning, reminiscent of reinforcement learning from human feedback (RLHF), could be eased with applications implementing linguistic methods:

> Giving explicitly the list of rules used to disambiguate a training corpus, rather than the corpus itself, would be more useful to the scientific community, as these rules could be examined, corrected, and maintained, while at the same time being applied to any text, and therefore be checked, falsified, and refined at will. In essence, this is what the linguistic approach proposes. (Silberztein, 2024, p. 21).

This contrasts stochastic AI models, since "when they produce inadequate results, they cannot be corrected or tweaked: the model must be rebuilt from scratch with new data" (Verdicchio, 2023, p. 12). The assumption that statistical approaches are less time-consuming than rule-based linguistic approaches thus reveals to be false (Silberztein, 2016, p. 22); the solution resides, again, in a better collaboration network that calls for the inclusion of computational linguists, who would be "more than happy" to help computer engineers with their expertise (Silberztein, 2024). In a very feasible view, that nonetheless seems rather utopian seen the state of the industry:

> Formalising the lexical, morphological, syntactic, and distributional properties of the standard English vocabulary would require the work of a dedicated team, much smaller than the gigantic groups assigned to the construction of tagged corpora for statistical NLP applications (statistical tagging or translation). A project like this would be beneficial for the whole linguistic community and would enable the development of NLP software with unequalled precision. (Silberztein, 2016, p. 27).

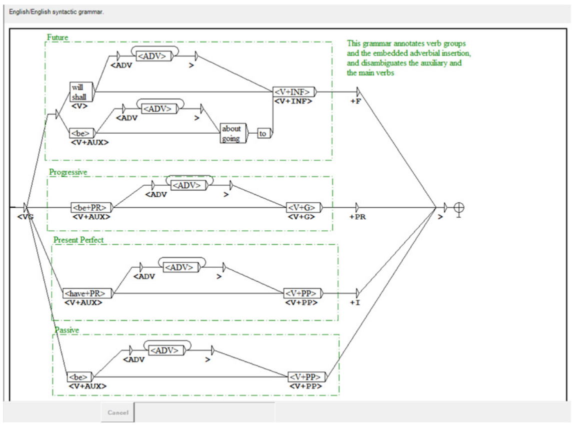
A hybrid approach would leverage both stochastic patterns and validated linguistic rules to produce more interpretable and reliable outputs. Such a methodology would also minimise dependency on large datasets, especially for low-resource languages or language nuances not well-represented in the available data, as well as for tasks that tackle out-of-domain specialistic language. Huang et al. (2020) already showed an example of combination of neural machine translation (NMT) with a rule-based layer made of dictionaries and formalised syntactic information, with promising results that encourage the integration of the latter to improve the performance of the former: especially in out-of-domain translations, their experiments showed that this hybrid approach can help in achieving higher precision. Also Wahde & Virgolin (2022) proposed a hybrid model for a conversational agent that is more transparent and interpretable thanks to a rule-based knowledge base, along with a set of principles (the five Is) that in their view governs transparency in AI: Interpretability, Inherent capability to explain, Independent data, Interactive learning, and Inquisitiveness (Wahde & Virgolin, 2022, pp. 1858-1859). It is from here that the approach presented in this paper is introduced, with a renewed focus on low-resource languages and the interest that linguists may show in its regard.

# 5 INTRODUCING NOOJ FOR HYBRID NLP

NooJ is here designated as a privileged choice for integrating the linguistic resources here discussed in future AI models. NooJ was introduced by the aforementioned scholar Max Silberztein (2004) and counts on a community in force of its being open source[4]. NooJ offers a powerful environment to develop resources that formalise linguistic phenomena, and to use them to parse texts and corpora, with NLF capacities that cover grammars of every level of the Chomsky's hierarchy, from Finite State Automata (FSA) to Turing machines in the form of Enhanced Recursive Transition Networks (ERTN), that while being half a century old demonstrates to be

[4] The *Association internationale des utilisateurs de NooJ* [International Association of NooJ Users] is an international scientific association and a non-profit organisation devoted to the advancement of linguistic research and CL. Its members consist of researchers, educators, linguists, and developers who collaborate to improve language processing techniques and support the integration of NooJ in both academic and professional settings. The association welcomes new members year-round and gathers every year since 2006 for an international conference dedicated to the various applications of the software (https://nooj.univ-fcomte.fr/association.html).

still relevant with new applications in the field (Chomsky & Schützenberger, 1963; Freidin, 2013).



English/English syntactic grammar.

This grammar annotates verb groups and the embedded adverbial insertion, and disambiguates the auxiliary and the main verbs

Figure 1: _VG.nog, a syntactic grammar that comes with the NooJ English module.

In ten years of history, contributions to CL using NooJ have been innumerable. Low-resource languages like Quechua (Duran, 2017; 2019) and Kabyle (Aoughlis et al., 2013) now have access to NLP thanks to NLF performed with the software. Moreover, with the use of NooJ, to cite a recent example, Walter Koza and María Mare (2023) were able to formalise resources that can not only recognise, but also generate texts with complex constructions involving the juxtaposition of many clitic pronouns in Spanish, while Mario Monteleone (2020) had showed how NooJ linguistic resources alone can be used to generate texts.

While it is true that NLF methodologies – in this case Maurice Gross' Lexicon-Grammar (Gross, 1994) – can be used on their own for such tasks, this paper proposes a methodology that integrates these rule-based systems into a possible AI model that draws its power from a deeper linguistic approach to leverage both structured linguistic rules and machine learning.

## 6 INTEGRATING NLF INTO AI: A PROPOSED METHODOLOGY

The advantage of this approach would lie in its dual focus: rule-based linguistic methods provide precision and control, while AI agents provide adaptability and generalisation. This synergy creates a more versatile and accurate language-comprehensive system that can be applied across various domains, from automated translation to educational tools: this model would offer a level of (linguistic) explainability lacking in pure AI systems, making it more reliable and easier to fine-tune (as lacunary performances would only require to touch specific resources like dictionaries and grammars, and not a whole model).

With a structured layer of linguistic rules, users can potentially track and explain why the model made certain linguistic choices while performing NLP tasks. For instance, in the context of text generation tasks, if a response is generated in a particular tense or with a specific syntactic structure, the model could reference the underlying linguistic rules it applied, making its decision process interpretable.

The integration of this additional layer involves NooJ's linguistic resources called in action to establish a solid linguistic foundation (see Figure 2). Reference corpora are piped into NooJ to be tokenised, annotated in electronic dictionaries and, in the same way, the syntactic and morphological structures present in such bodies of data are parsed by finite-state grammars. In this scenario, the rule-based framework serves as a guide that constrains and informs the AI model, ensuring that its outputs adhere closely to established language norms.

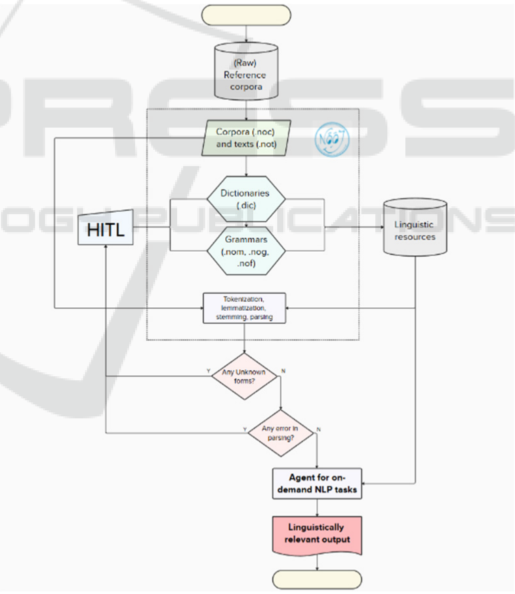


Figure 2: Flowchart for the proposed methodology.

To implement this methodology, a pipeline could be developed where NooJ's resources first preprocess the input textual data, structuring it in a way that aligns with linguistic rules formalised thanks to NooJ's powerful linguistic engine. These structured inputs are then fed into a machine learning model trained to recognise and generate text that aligns with this structure. The model could use techniques such as transformers, which have a strong ability to capture long-range dependencies and context in text; in any case, this proposition doesn’t aim to propose any particular instruction of such final step, and focuses rather on the layer to be developed looking at linguistically-informed ways of processing input data: morphological, syntactic, and semantic parsing are introduced as a pre-processing layer before on-demand NLP tasks (such as text generation, spell-checking and machine translation), with such layer susceptible of the intervention of skilled linguists, the humans in the loop (HITL).

By grounding these models in structured linguistic data, developers would give them a robust linguistic framework to work within, which helps them not only to produce coherent and linguistically accurate text, but also to make such output the result of an operation that involves a “linguistic” rather than merely “mathematical” intelligence.

## 7 CONCLUSIONS

This paper, oriented to a better future for NLP, proposes a framework in which linguistic knowledge participates as the main actor in the creation of agents that can achieve actual goals in terms of linguistic competence, aligning more closely with human cognitive processes and thus contributing to a significant qualitative step in the very definition of Artificial Intelligence, while also aligning more closely with the goals and needs of computational linguistics.

Modern NLP is indeed powered by machine learning, but linguistic theories and methodologies should continue to deeply inform its development. NLP will be far more effective without linguistics guiding how we as humans understand, structure, and interpret language: here’s the reason why to address an interdisciplinary effort that bridges a gap that is now the elephant in the room for language scholars interested in digital methodologies. AI researchers and engineers should collaborate closely with linguists, ensuring that models respect the complexities of language structures and phenomena. Such an approach would pave the way for AI systems capable of genuinely engaging in natural, human-like communication, enhancing their utility and trustworthiness across a wide range of applications; the first of which is the possibility to retrieve the linguistic reasoning behind AI’s manipulation of textual data, paving the way for a future in which low-resource languages will see a degree of representation that allows for the engagement in the development and utilisation of digital resources.

NooJ complex linguistic resources could be embedded directly into the AI pipeline through command-line calls (noojapply.exe), acting as pre-processing for NLP tasks. Moreover, if NooJ resources are exposed through APIs, they can be called to perform real-time checks as part of an AI application, improving real-time analytical accuracy of the model’s performance, so to finally bring back NL together with P in a way that doesn’t only strive for some semblable aesthetic imitation of the products of human language, but that first and foremost cares about the linguistic intelligence that can (and, in my view, must) be made artificial for such purpose. With such intelligence comes a renewed ethical framework in approaching a vaster number of low-resource languages, and with such ethical framework comes a more sustainable future workflow.

## ACKNOWLEDGEMENTS

Special thanks to Giuseppe Laquidara, who revised an early version of this paper and helped me perfect it with his invaluable advice.

## REFERENCES

Aoughlis, F., Naït-Zerrad, K., Annouz, H., Kaci, F., & Habet, M. S. (2014). A New Tamazight Module for NooJ. In Koeva, S., Mesfar, S., & Silberztein, M. (eds.), *Formalising Natural Languages with NooJ 2013: Selected papers from the NooJ 2013 International Conference*, pp. 13-21. Cambridge Scholars Publishing.

Barriga Martínez, D., & Mijangos, V., Gutierrez-Vasques, X. (2021). Automatic Interlinear Glossing for Otomi Language. In Mager, M., Oncevay, A., Rios, A., Meza Ruiz, I. V., Palmer, A., Neubig, G., & Kann, K. (eds.), *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pp. 34–43. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.americasnlp-1.5.

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In Jurafsky, D., Chai, J., Schluter, N., & Tetreault, J. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5185–5198. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463.

Bender, E. M., Gebru, T.; McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21),* pp. 610–623. Association for Computing Machinery. https://dl.acm.org/doi/10.1145/3442188.3445922.

Chomsky, N, & Schützenberger, M. P. (1963). The algebraic theory of context free languages. In Braffort, P., & Hirschberg, D. (eds.), *Computer Programming and Formal Systems*, pp. 118–161. North Holland.

Duran M. (2017). Quechua Module for NooJ Multilingual Linguistic Resources for MT. In Barone, L., Monteleone, M., Silberztein, M. (eds.), *Automatic Processing of Natural Language Electronic Texts with NooJ: Selected Papers from the International Conference NooJ 2016*, pp. 48-63. Springer.

Duran, M. (2019). Can a robot help save an endangered language? In Adda, G., Choukri, K., Kasinskaite-Buddeberg, I., Mariani, J., Mazo, H., & Sakriani, S. (eds.), *Proceedings of the Language Technologies for All*, pp. 284-287. UNESCO.

Fanni, S.C., Febi, M., Aghakhanyan, G., Neri, E. (2023). Natural Language Processing. In Klontzas, M.E., Fanni, S.C., Neri, E. (eds.), *Introduction to Artificial Intelligence. Imaging Informatics for Healthcare Professionals* (pp. 87-99). Springer, Cham. https://doi.org/10.1007/978-3-031-25928-9_5.

Freidin, R. (2013). Noam Chomsky's contribution to linguistics. In Allan, K. (ed.), *The Oxford Handbook of the History of Linguistics* (pp. 438-467). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199585847.013.0020.

Gross, Maurice (1994). Constructing Lexicon-Grammars. In Beryl T., Atkins, S. & Zampolli, A. (eds.), *Computational approaches to the lexicon*, 213-263. Oxford OUP.

Huang, J., Lee, K., & Kim, Y. (2020). Hybrid translation with classification: Revisiting rule-based and neural machine translation. *Electronics*, *9*(2), 201. https://doi.org/10.3390/electronics9020201.

Jiang, M., Lin, B. Y., Wang, S., Xu, Y., Yu, W., & Zhu, C. (2023). *Knowledge-augmented methods for natural language processing*. Springer.

Koza, W., & Mare, M. (2023). Computational modeling of a generative grammar for clitic order in River Plate Spanish. *Revista Española de Lingüística Aplicada/Spanish Journal of Applied Linguistics*, *36*(1), 270–306. https://doi.org/10.1075/resla.20053.koz.

Lewis, J. E., Abdilla, A., Arista, N., Baker, K., Benesiinaabandan, S., Brown, M., Cheung, M., Coleman, M., Cordes, A., Davison, J., Duncan, K., Garzon, S., Harrell, D. F., Jones, P. L., Kealiikanakaoleohaililani, K., Kelleher, M., Kite, S., Lagon, O., Leigh, J., Levesque, M., Mahelona, K., Moses, C., Nahuewai, I., Noe, K., Olson, D., Parker Jones, 'Ō., Running Wolf, C., Running Wolf, M., Silva, M., Fragnito, S., & Whaanga, H. (2020). Indigenous Protocol and Artificial Intelligence Position Paper. Project Report. *Indigenous Protocol and Artificial Intelligence Working Group and the Canadian Institute for Advanced Research*, Honolulu, HI.

McMillan-Major, A. (2020). Automating Gloss Generation in Interlinear Glossed Text. *Society for Computation in Linguistics*, *3*(1), 338-349. https://doi.org/10.7275/tsmk-sa32.

Moeller, S., Liu, L., Yang, C., Kann, K., & Hulden, M. (2020). IGT2P: From Interlinear Glossed Texts to Paradigms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5251–5262. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.424.

Monteleone, M. (2020). Automatic Text Generation: How to Write the Plot of a Novel with NooJ. In Fehri, H., Mesfar, S. & Silberztein, M. (eds.), *Formalizing Natural Languages with NooJ 2019 and Its Natural Language Processing Applications 13th International Conference, NooJ 2019*, Hammamet, Tunisia, June 7–9, 2019 (pp. 135-146). Springer.

Pillai, A. S., & Tedesco, R. (2024). *Machine learning and deep learning in natural language processing*. CRC Press.

Sampson, G. (2002). *Empirical linguistics*. A&C Black.

Schütze, C. T. (2016 [1996]). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Language Science Press.

Silberztein, M. (2004). NooJ: A Cooperative, Object-Oriented Architecture for NLP. In *INTEX pour la Linguistique et le traitement automatique des langues*. Cahiers de la MSH Ledoux, Presses Universitaires de Franche-Comté. https://books.openedition.org/pufc/30117.

Silberztein, M. (2016). *Formalizing natural languages: The NooJ approach*. John Wiley & Sons.

Silberztein, M. (2024). Preface. In Silberztein, M. (ed.), *Linguistic Resources for Natural Language Processing* (pp. xviii-xvi). Springer Nature.

Verdicchio, M. (2023). Che cosa genera davvero l'IA generativa? *ParadoXa*, *7*(4), 29-43.

Wahde, M., & Virgolin, M. (2022). DAISY: An Implementation of Five Core Principles for Transparent and Accountable Conversational AI. *International Journal of Human–Computer Interaction*, *39*(9), 1856–1873. https://doi.org/10.1080/10447318.2022.2081762.

Yogatama, D., de Masson d'Autume, C., Connor, J., Kocisky, T., Chrzanowski, M., Kong, L., Lazaridou, A., Ling, W., Yu, L., Dyer, C., & Blunsom, P. (2019). *Learning and evaluating general linguistic intelligence*. arXiv. https://arxiv.org/abs/1901.11373.

Zhao, X., Ozaki, S., Anastasopoulos, A., Neubig, G., & Levin, L. (2020). Automatic Interlinear Glossing for Under-Resourced Languages Leveraging Translations. In Scott, D., Bel, N., & Zong, C. (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 5397–5408. International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.471.

Zianouka, Y., Hetsevich, Y., Latyshevich, D., & Dzenisiuk, Z. (2022). Automatic Generation of Intonation Marks and Prosodic Segmentation for Belarusian NooJ Module. In Bigey, M., Richeton, A., Silberztein, M., Thomas, I. (eds.), *Formalizing Natural Languages: Applications to Natural Language Processing and Digital Humanities. 15th International Conference* (pp. 231-242)*,* Besançon, France, June 9–11, 2021. Springer.