Grammar Assistance Using Syntactic Structures (GAUSS)

Olga Zamaraeva[1], Lorena S. Allegue[2], Carlos Gómez-Rodríguez[1], Margarita Alonso-Ramos[2] and Anastasiia Ogneva[3]

[1]Universidade da Coruña, CITIC, Department of Computer Science and Information Technologies. 15071 A Coruña, Spain
[2]Universidade da Coruña, CITIC, Department of Humanities ("Letras"). 15071 A Coruña, Spain
[3]Universidade de Santiago de Compostela, Department of Developmental Psychology, 15782 Santiago de Compostela, Spain

Abstract

Automatic grammar coaching serves an important purpose of advising on standard grammar varieties while not imposing social pressures or reinforcing established social roles. Such systems already exist but most of them are for English and few of them offer meaningful feedback. Furthermore, they typically rely completely on neural methods and require huge computational resources which most of the world cannot afford. We propose a grammar coaching system for Spanish that relies on (i) a rich linguistic formalism capable of giving informative feedback; and (ii) a faster parsing algorithm which makes using this formalism practical in a real-world application. The approach is feasible for any language for which there is a computerized grammar and is less reliant on expensive and environmentally costly neural methods. We seek to contribute to Greener AI and to address global education challenges by raising the standards of inclusivity and engagement in grammar coaching.

Keywords

grammar engineering, grammar coaching, second language acquisition, HPSG, syntactic theory, syntax, parsing

1. Introduction

The GAUSS project is concerned with **a new, faster parsing technology for grammar coaching** and will develop a Spanish grammar coaching system. Automatic grammar coaching helps people write more like a native speaker of a language would, thus helping them navigate around biases associated with language. This is important for (i) finding a job and counterbalancing latent discrimination in any given society, in the case of major languages like Spanish; and (ii) reinforcing the understanding that each language has a systematic grammar in its own right, in the case of minority languages (like e.g. Galician). Grammar coaching systems rely on parsing to determine (i) that grammar in a sentence could be improved; and (ii) how specifically to improve it. Parsing is mapping a sentence to a structure (Figure 1).

The project uses an implemented linguistic grammar of Spanish to provide meaningful feedback on writing.

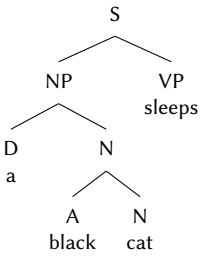

Figure 1: A simplified syntactic parse.

The notion of grammaticality encoded in such grammars is more descriptive than prescriptive; the system will not try reinforce someone's opinion on what is correct and what is not. Our specific contribution will be in (i) developing such a system for Spanish, one of the world's most spoken languages, leveraging an existing body of linguistic knowledge; and (ii) making the underlying parsing technology fast enough to be deployed at scale. A Spanish system based on cross-linguistically applicable methodology will pave the way for other European languages including minority languages, starting from Galician, the language of our host province. The main challenge we will address is integrating neural and symbolic approaches to parsing, demonstrating that expensive neural methods can be applied in a limited manner, and that the computational "price tag" of NLP technology can be reduced.

SEPLN-CEDI-PD 2024: Seminar of the Spanish Society for Natural Language Processing: Projects and Demonstrations, June 19-20, 2024, A Coruña, Spain

olga.zamaraeva@udc.es (O. Zamaraeva); l.sallegue@udc.es (L. S. Allegue); carlos.gomez@udc.es (C. Gómez-Rodríguez); margarita.alonso@udc.es (M. Alonso-Ramos); anastasiia.ogneva@usc.es (A. Ogneva)
https://olzama.github.io/ (O. Zamaraeva); https://www.grupolys.org/~cgomez (C. Gómez-Rodríguez); https://sites.google.com/view/anastasiiaogneva (A. Ogneva)
0000-0001-9969-058X (O. Zamaraeva); 0009-0009-5529-4150 (L. S. Allegue); 0000-0003-0752-8812 (C. Gómez-Rodríguez); 0000-0002-1353-9270 (M. Alonso-Ramos); 0000-0003-0237-7146 (A. Ogneva)

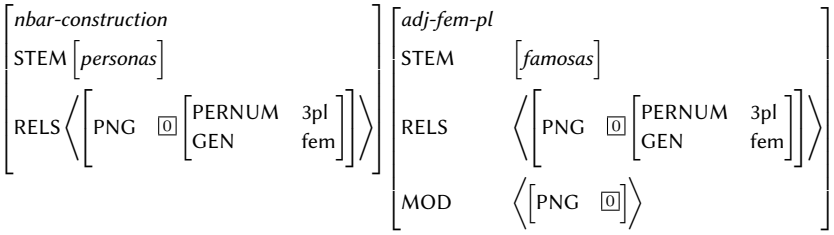

Figure 2: Two simplified HPSG structures that can form a phrase 'famous persons' in Spanish, *personas famosas*.

2. State of the art at the start of the project

Most grammar coaching systems available today are purely statistical and do not use explicit linguistic knowledge. Based on purely statistical methods and lacking interpretability, they "guess" based on the context and are not aware of concepts like agreement. Their feedback is divorced from the methodology of suggesting a better sentence, opening possibilities for wrong feedback. Such systems are often only available for English, because their neural architectures require huge quantities of training data. Such systems are also ecologically problematic[1].

3. Methodology

The GAUSS project is the result of the collaboration between research areas such as CS, NLP, theoretical linguistics, and applied linguistics. The intersectional nature of the project is realized by the combination of NLP techniques and theoretically formalized grammars. In particular, the project relies on the Spanish Resource Grammar [SRG; 2, 3, 4], a grammar of Spanish implemented in the Head-driven Phrase Structure Grammar formalism (HPSG).

3.1. HPSG syntax theory

Head-driven Phrase Structure Grammar [HPSG; 5] is a constraint unification theory of syntax. A sentence is analyzed as a structure where parts can be constrained to be identical to each other. For example, a verb's agreement values (e.g. third person) can be constrained to be identical to the agreement values of the subject of the verb. Similarly, adjectives can be constrained with respect to the agreement values of the noun they modify, as shown in Figure 2. Crucially, ungrammatical strings of words will violate the constraints required for well-formed structures and as such will not be covered by an HPSG grammar.

Structures like the ones in Figure 2 are instances of more general types and can be seen in the specific results of deploying the grammar on some data. The grammar itself contains the types, not the instances. The types are instantiated through interfacing with the lexicon and, in some cases, an external morphophonological analyzer.

The HPSG theory covers many syntactic phenomena and has been developed and tested using a variety of data from a variety of languages. One of the approaches to the empirical testing of this theory is implementing it on the computer and then automatically parsing data and inspecting the results for correctness and consistency. Efforts of this kind include ParGram [6], CoreGram [7] and DELPH-IN [8, 9] It is this approach that gave rise to the SRG.

3.2. DELPH-IN Consortium

The DELPH-IN research consortium is an international effort for grammar engineering using HPSG: Deep Linguistic Processing with HPSG Initiative. It is committed to using a particular version of the HPSG formalism that was defined originally in [8]. The consortium develops tools such as parsers, including the parser we used in this project, the ACE parser [10]. Another set of relevant tools includes the software for automatic profiling of test data known as `incr tsdb()` (pronounced 'tsdb++') [11, 12] and a related tool "full-forest treebanker" (fftb) [13]. These tools allow us to inspect differences between different grammar versions systematically.

Grammars are tested on sentences automatically, using a parser. The first time a grammar is run on a sentence, an expert must verify the correctness of the output. Often it makes sense to do this by looking at the semantic (dependency) structure; we can assume that if the semantics is correct, then the syntactic structure that corresponds to it is adequate. The semantics in DELPH-IN grammars is modeled with Minimal Recursion Semantics formalism [MRS; 14]. An MRS structure is a bag of predications encoding dependencies as well as modifier and negation scope, information structure, and more. It can be automatically converted to a dependency structure familiar to natural language processing (NLP) practitioners (Figure 3). When the parser analyzes a sentence according to the grammar, the resulting structure includes an

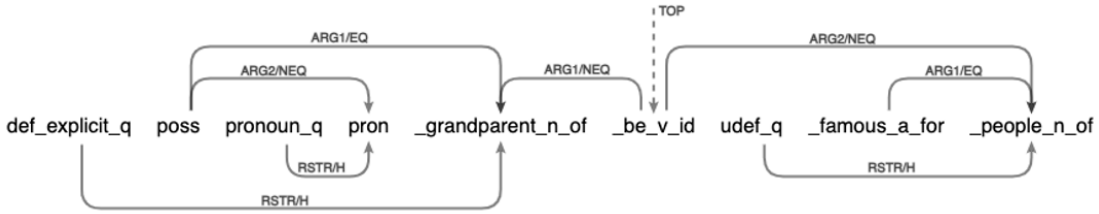

Figure 3: Dependency structure for "My grandparents are famous people".

MRS, the adequacy of which is easy to establish manually (whether the meaning of the sentence is the intended one). Adequacy of obtained analyses on corpora serve as accumulating evidence for the validity of the theory of syntax.

3.3. Spanish Resource Grammar

At the core of the project's methodology is the digital representation of the Spanish syntax, the Spanish Resource Grammar [2, 3, 4]. The SRG consists of 54,510 lemmas in the lexicon, 543 lexical types to instantiate those lemmas, 504 lexical rule types serving morphophonological analysis, and 226 phrasal types. It is the second largest DELPH-IN grammar (after the English Resource Grammar [15, 16]). SRG was first developed prior to the ACE parser and one of the objectives of the GAUSS project ended up being the complete reimplementation of the SRG morphophonological interface. The outcome is that the SRG can now be used with the ACE parser [4]. As before, it relies on an external morphophonological analyzer Freeling [17].

One major outcome of this is that we could reparse the portions of the AnCora corpus previously released as the TIBIDABO treebank [3]. The previously released version was partially verified for the correctness of the structure but the accuracy figures corresponding to that verification were never reported (as far as we can tell). One of the outcomes of GAUSS is the re-parserd, re-verified, and re-released portions of TIBIDABO (currently 2291 sentences) [4]. The updated version of the SRG along with the verified treebanks are open-source and are released on GitHub: https://github.com/delph-in/srg

3.4. Using the SRG with learner data

The main idea behind the GAUSS project is that we can use the SRG to model constructions characteristic of learners of Spanish (as opposed to native speakers). We create a version of the SRG that is modified specifically to cover learner constructions, starting with gender agreement constructions, like the one illustrated in example (1).

(1) *Mis abuelos son
my.3PL grandparent.MASC.PL be.3PL.PRES.IND
personas famosos.
person.FEM.3PL famous.MASC.PL
Intended: 'My grandparents are famous people.'
[spa; Yamada et al. 18]

The grammar will detect such learner structures using what is called 'mal-rules' [19], a technical term for HPSG types designed specifically to cover productions characteristic of learners. For example, the grammar will have to have a way to ignore the incompatible agreement values in Figure 4.

We achieve this by only a small set of modifications to the grammar. We use the interface of the grammar with the external morphophonological analyzer to recognize any noun or adjective as potentially belonging to either gender (this requires 40 short additional entries in the lexical rule section of the grammar, one corresponding to each possible Freeling noun or adjective tag). We associate each such lexical rule with a special LEARNER feature, so that ultimately any sentence that uses one or more of such rules can be detected as a learner production. No changes in the syntax part of the grammar are required, in principle. However, deploying the grammar on the learner sentences without modifications revealed a number of overgeneration issues in the original grammar, which we were able to fix thanks to this experiment. Overgeneration is when a grammar covers an ungrammatical sentence or produces a nonsensical structure for a sentence along with the correct one(s). When we saw instances of the original grammar covering learner productions, we investigated such cases and have found 4 syntactic types (so far) which were underconstrained with respect to the agreement values. We have added the missing agreement constraints, which resulted in reduced overgeneration and ambiguity of the SRG with respect to the TIBIDABO treebank. In this way, modeling learner constructions helped us improve the analysis of agreement in the original SRG.

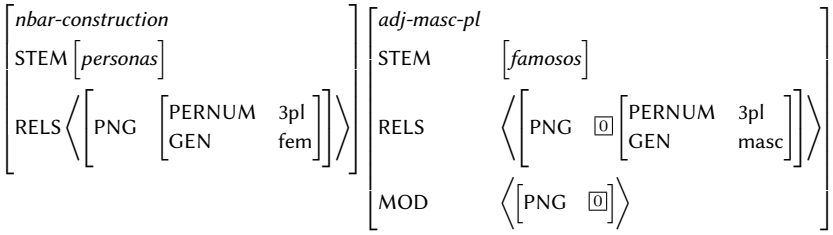

Figure 4: HPSG feature structures representing incompatible agreement values in the learner production *personas famosos*.

After all the necessary mal-rules are implemented, the plan is to (1) accompany each model of a learner construction with meaningful feedback; and (2) deploy the grammar as a web-based service such that it can be tested by learners of Spanish. This is work in progress.

3.5. Parsing speed bottleneck

The main challenge in HPSG parsing speed is that large feature structures combinatorically lead to huge search space. As a result, HPSG parsing is comparatively slow in practice. For example, the ACE parser takes about 3 seconds per sentence on average on a corpus of 100K sentences (some of these sentences take minutes while others take less than a second) [20]. The GAUSS project attempts to address this challenge by a combination of methodologies: (1) improving analyses in the grammar to reduce meaningless ambiguity (overgeneration) and thus reduce the size of the parse chart; (2) integrating top-down parsing, and (3) filtering lexical entries and grammar rules so that fewer rules are considered at each step. Method (1) is what we employed while addressing overgeneration we discovered by deploying the grammar on the learner corpus. We have managed to improve the SRG's performance up to 60% on sentences of length 8-10. Method (2) has been underexplored in HPSG but has seen a rekindled interest recently [21]. HPSG parers are overwhelmingly bottom-up but for long sentences, a lot can be learned immediately from the start of the sentence/top of the syntax tree, discarding many irrelevant search paths. Method (3) includes developing a neural supertagger (filter) for HPSG. The supertagger will reduce the number of possibilities the parser needs to explore by discarding unlikely word meanings. Statistical filtering was successfully applied to HPSG [22], and we are now researching how neural methods can improve the SOTA. We start with applying method (3) to the English Resource Grammar treebanks and obtain a speed-up of a factor of three compared to the baseline. However, when we attempted the method on the Spanish treebanks, the results were not yet satisfactory, apparently because the Spanish treebanks were not big enough at the start of the GAUSS project. Now that we added more verified items in the treebanks, we can attempt to train a neural supertagger for Spanish once again.

4. Planning and Team

The GAUSS project consists of three Research Objective (RO) and four Work Packages (WP). They are summarized in Table 1.

Table 1
Research Objectives (RO)

RO	WP	Objective
RO1	WP1	Fast HPSG parsing
RO2	WP2	Spanish error productions in HPSG
RO3	WP3–4	Empirical integration of RO1–2

The team consists of the PI MSCA postdoctoral fellow Olga Zamaraeva, supervisor Carlos Gómez-Rodríguez, co-supervisor Margarita Alonso-Ramos, collaborator Anastasiia Ogneva, and research assistant Lorena S. Allegue. Olga Zamaraeva does most of the technical and organizational work. Lorena S. Allegue verifies the correctness of the grammar output. Carlos Gómez-Rodríguez advises on computational issues. Margarita Alonso-Ramos advises on the use of the learner corpora. Anastasiia Ogneva advises on second language acquisition theory.

Acknowledgments

The GAUSS project is funded by the European Union's Horizon Europe Framework Programme under the Marie Skłodowska-Curie postdoctoral fellowship grant HORIZON-MSCA-2021-PF-01 (GAUSS, grant agreement No 101063104) The project is carried out in the Language and Society Information research group (LyS) of Universidade da Coruña.

References

[1] R. Schwartz, J. Dodge, N. A. Smith, O. Etzioni, Green AI, ACM 63 (2020).

[2] M. Marimon, The Spanish Resource Grammar, in: LREC, 2010.
[3] M. Marimon, N. Bel, L. Padró, Automatic selection of HPSG-parsed sentences for treebank construction, Computational Linguistics 40 (2014) 523–531.
[4] O. Zamaraeva, L. S. Allegue, C. Gómez-Rodríguez, Spanish Resource Grammar version 2023, in: COLING-2024, in press.
[5] C. Pollard, I. Sag, Head-Driven Phrase Structure Grammar, CSLI, 1994.
[6] M. Butt, T. H. King, Urdu and the parallel grammar project, in: Proceedings of the 3rd workshop on Asian language resources and international standardization-Volume 12, Association for Computational Linguistics, 2002, pp. 1–3.
[7] S. Müller, The CoreGram project: Theoretical linguistics, theory development and verification, Journal of Language Modelling 3 (2015) 21–86.
[8] A. Copestake, Appendix: Definitions of typed feature structures, Natural Language Engineering 6 (2000) 109–112.
[9] E. M. Bender, G. Emerson, Computational linguistics and grammar engineering, in: S. Müller, A. Abeillé, R. D. Borsley, J.-P. Koenig (Eds.), Head-Driven Phrase Structure Grammar: The handbook, 2021.
[10] B. Crysmann, W. Packard, Towards efficient HPSG generation for German, a non-configurational language., in: COLING, 2012, pp. 695–710.
[11] S. Oepen, D. Flickinger, Towards systematic grammar profiling. test suite technology 10 years after, Computer Speech & Language 12 (1998) 411–435.
[12] S. Oepen, [incr tsdb ()] competence and performance laboratory. user and reference manual, 1999.
[13] W. Packard, UW-MRS: Leveraging a deep grammar for robotic spatial commands, SemEval 2014 (2014) 812.
[14] A. Copestake, D. Flickinger, C. Pollard, I. A. Sag, Minimal recursion semantics: An introduction, Research on language and computation 3 (2005) 281–332.
[15] D. Flickinger, On building a more efficient grammar by exploiting types, Natural Language Engineering 6 (2000) 15–28.
[16] D. Flickinger, Accuracy v. robustness in grammar engineering, in: E. M. Bender, J. E. Arnold (Eds.), Language from a Cognitive Perspective: Grammar, Usage and Processing, CSLI, Stanford, CA, 2011, pp. 31–50.
[17] X. Carreras, I. Chao, L. Padró, M. Padró, Freeling: An open-source suite of language analyzers, in: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04), 2004.
[18] A. Yamada, S. Davidson, P. Fernández-Mira, A. Carando, K. Sagae, C. Sánchez-Gutiérrez, COWS-L2H: A corpus of Spanish learner writing, Research in Corpus Linguistics 8 (2020) 17–32.
[19] D. Schneider, K. McCoy, Recognizing syntactic errors in the writing of second language learners, in: ACL, 1998, pp. 1198–1204.
[20] O. Zamaraeva, C. Gómez-Rodríguez, Revisiting supertagging for HPSG, 2023. arXiv:2309.07590.
[21] L. Chiruzzo, D. Wonsever, Statistical deep parsing for spanish using neural networks, in: IWPT, 2020, pp. 132–144.
[22] R. Dridan, Ubertagging: Joint segmentation and supertagging for english, in: EMNLP, 2013, pp. 1201–1212.