Automatic Treatment of Causal, Consecutive, and Counterargumentative Discourse Connectors in Spanish

A Pedagogical Application of NooJ

Andrea Rodrigo[1(✉)], Silvia Reyes[1], and María Andrea Fernández Gallino[2]

[1] Facultad de Humanidades y Artes, Universidad Nacional de Rosario, Rosario, Argentina
andreafrodrigo@yahoo.com.ar, sisureyes@gmail.com
[2] Facultad Regional Rosario, Universidad Tecnológica Nacional, Rosario, Argentina
mandrea.fernandezg@gmail.com

Abstract. Our intention, within the framework of the pedagogical application of NooJ to Spanish teaching undertaken by the research team Argentina (Centro de Estudios de Tecnología Educativa y Herramientas Informáticas, UNR), is to continue to apply discourse tags, as we have done in previous work dealing with causal discourse connectors in Spanish. We added counterargumentative and consecutive discourse connectors to our dictionary, taking into consideration a specific population of Spanish learners whose mother tongue is Spanish, but who sometimes face the same difficulties as those experienced by learners of Spanish as a foreign language (SFL). The corpus comprises stories for kids written by students of two tertiary colleges for primary education teachers. The corpus texts show enormous deficits. One of the main deficiencies lies in the lack of resources enabling metalinguistic reflection, which in turn, in our teaching experience, is the motor of linguistic knowledge. Another difficulty tangentially lies in a lack of clarity concerning the correction and self-correction of texts, since there is a belief that writing is the result of inspiration and that it is done in a flash. NooJ can give at least a partial answer to these issues, because dictionaries and grammars contribute to the formalization of linguistic knowledge. Our idea is to show some syntactic grammars created with the NooJ platform in order to analyse corpus phrases and make visible the use of causal, consecutive, or counterargumentative connectors.

Keywords: NLP · Pedagogy · Causal connectors · Consecutive connectors · Counterargumentative connectors · Spanish language · NooJ

1 Introduction

1.1 The Purpose of Our Paper

Within the framework of the pedagogical application of NooJ to Spanish teaching undertaken by the research team Argentina (Centro de Estudios de Tecnología Educativa y

B. Bekavac et al. (Eds.): NooJ 2020, CCIS 1389, pp. 159–169, 2021.
https://doi.org/10.1007/978-3-030-70629-6_14

Herramientas Informáticas, UNR, Argentina), we first focused on working with grammatical categories such as nouns, adjectives, and adverbs. However, some difficulties in learning were noted in native speakers of Spanish, since linguistic terminology is not always fully understood by them and on occasions it may become an obstacle in itself. Sometimes these difficulties arise because linguistic knowledge acquired by them long ago in language classes has been forgotten. Consequently, we turned our attention to discourse tags because the relations of cause, consequence, and counter argumentation are universal, and thus it is easier to have recourse to that knowledge.

In line with these ideas, we made progress towards causal discourse connector tags, thinking of learners whose mother tongue is Portuguese [6]. On this basis, we inserted this type of connector in our NooJ dictionaries. Our intention in this paper has been to continue with the introduction of discourse tags by adding counterargumentative and consecutive connectors to our dictionary. Unlike previous papers, in this opportunity we take as reference a population of students of two tertiary colleges for primary education teachers whose average age ranges from 18 to 25 years. They constitute a heterogeneous group of low socioeconomic status who in fact study with a view to finding a job after graduation. These students generally completed their secondary studies with interruptions, or even started other careers which they had to abandon due to financial or family problems. We chose this specific population because we considered it ideal for us to enrich our research with a new perspective. Although these learners are native speakers whose mother tongue is Spanish, they sometimes face the same difficulties as those experienced by learners of Spanish as a foreign language (SFL). Disjointed sentences and decidedly complex syntax are a constant. For this reason, our contribution using NooJ is relevant for addressing discourse connectors.

1.2 The Research Subject Matter: Discourse Connectors

To conceptualize connectors we relied on Martín Zorraquino and Portolés, who define a connector as "*a discourse marker that semantically and pragmatically relates a discourse member with another discourse member*" [1 p. 4093]. The assembly effect resulting from the presence of discourse connectors precisely explains the fact that we do not have the feeling that we are reading disconnected phrases. Otherwise, if there are no discourse connectors, the understanding of the relationship between phrases is left to the reader, who must deduce the cohesion between them.

In this paper, we particularly address three types of connectors: causal, counterargumentative, and consecutive connectors. It is interesting to highlight what Montolío [2] states about causal connectors, for they imply "*a kind of instruction given to the interlocutor, of the type 'what follows constitutes the cause of what has been said before' (meant, for example, by porque*", or how she describes consecutive connectors: "*what follows is the conclusion that is deduced from previous information*" [2 p. 29]. Finally, with respect to counterargumentative connectors, Quintero Ramírez [4] defines them as those that "*establish a relationship in which what is represented in the second textual segment contrasts or limits what is expressed in the first segment*" [4 p. 51]. As an additional detail, it is worth noting that one of the most common counterargumentative connectors in Spanish is *pero* 'but', whose main function is to divert "*the argumentative line of the previous sentence*" [2 p. 23].

1.3 The Corpus

The writing of children's tales was undoubtedly a major challenge for our tertiary students because of the difficulties we have already stated above, which are centred on sometimes discontinuous or fragmented secondary education training. However, the aim of the *Workshop on Text Comprehension and Production* was to bring into play essential textual skills for their future work as primary education teachers.[1] In this respect, after doing revision of the basic notions of syntax and semantics in Spanish, students were asked to write a tale. The only requirement was that the story should be addressed to children who are eight to twelve years old. Fortunately, students voluntarily signed a consent form so that their texts could be analysed anonymously by our research team.

Generally speaking, the texts show enormous deficits. One of the main deficiencies is the lack of resources enabling metalinguistic reflection, which in turn, in our teaching experience, is the motor of linguistic knowledge. Another difficulty tangentially lies in a certain lack of clarity concerning the correction and self-correction of texts, as there is a belief that writing is the result of inspiration and that it is done in a flash.

The corpus of short stories comprises 49,462 characters and 8,927 words. As regards orthography, many of these words present spelling difficulties, and it would be possible to create new dictionaries to include those words. As a general rule, the writing lacks accents, punctuation marks, capitalization, and proper spelling. For example, in one instance, the word *nietos* 'grandchildren' was misspelt as *ñetos*. We are aware that the sample we are analysing only constitutes a small part and that we cannot jump to major conclusions until the sample is enlarged. However, we are in a position to offer a useful appreciation of the situation. In our opinion, the reasons for this carelessness in writing have to do, more often than not, with mobile phone textspeak and chats, whose immediacy hinders the correction of errors. What Odrowąż-Coates [3] states about the English language can apply to other languages such as Spanish:

> *The popularity of English is reinforced online not only by software makers, but also by Internet users who embrace a new spelling code for English deemed as cool and hip [cut]. It has no formal corpus and is orthographically flexible. It grew in popularity due to netspeak or chat speak, which is an interaction with the use of live text online. The IT-mediated code used by users in electronic mailing, instant messaging, SMS, or chat, known as textese or textism is also based on English.* [3 p.85]

[1] These students are being trained to become primary education teachers at two tertiary colleges: *Escuela Normal Superior* N°35 "*J. M. Gutiérrez*" and *Escuela Normal Superior* N°36, "*M. Moreno*". We acknowledge Camila Ferramondo for compiling the corpus.

As we all can see, we are witnessing changes in writing which are the result of the irruption of information technology into everyday life. In our view, there is no point in making judgements about these changes because they should be understood in a more profound sense and not only "standardized". This is the fundamental premise that forms the basis of our work.

2 Working with NooJ

2.1 Why NooJ?

The tool developed by Silberztein [7, 8] constitutes a useful instrument to analyse the reference corpus, since it allows us to validate whether our linguistic descriptions are accurate or not. The pedagogical application of NooJ is founded on the metalinguistic reflection that arises during the process.

> *When linguistic descriptions have been entered into a computer, a computer can apply them to very large texts in order to extract from these texts examples or counterexamples that validate (or not) these descriptions* [cut]. *Finally, the description of certain linguistic phenomena makes it possible to construct NLP software applications.* [8 p. 6]

NooJ has indeed the advantage of not being a black box, and accordingly, new linguistic information can always be entered into the dictionaries and grammars created by researchers in order to process and systematize students' written texts as regards, for example, accentuation, punctuation, capitalization, and spelling, which are the aspects that attracted our attention in the previous section.

2.2 The Dictionary of Discourse Connectors

With the purpose of formalizing the Spanish expressions containing discourse connectors, new tags were introduced into the Spanish Module Argentina dictionaries [9]. In a previous study, we introduced a new tag into the Properties´ Definition file by incorporating a new category to name discourse connectors, Connector [C]. And within this category, the tag [+caus] was added to name the subclass of causal discourse connectors: [C+caus] [6]. In the present study, we added two new subclasses of discourse connectors [C+consec] and [C+contrarg], which refer to consecutive and counterargumentative discourse connectors, respectively.

Today, our dictionary includes some of the most usual causal, consecutive, and counterargumentative discourse connectors (Fig. 1).

```
# Special Characters: '\' '"' ' ' ',' '+' '-' '#'
#

porque,C+caus
gracias a,C+caus
gracias al,C+caus
ya que,C+caus
a causa de,C+caus
por eso,C+caus
pues,C+caus
puesto que,C+caus
dado que,C+caus
por el hecho de que,C+caus
en virtud de,C+caus
por tanto,C+consec
en consecuencia,C+consec
por consiguiente,C+consec
de ahí,C+consec
así pues,C+consec
pero,C+contrarg
en cambio,C+contrarg
por el contrario,C+contrarg
sin embargo,C+contrarg
no obstante,C+contrarg
con todo,C+contrarg
ahora bien,C+contrarg
```

Fig. 1. Excerpt from the dictionary of connectors in Spanish

The next step was to collect information about the connectors in our corpus. To do so, we first apply ***TEXT > Locate*** to search for discourse connectors <C> in the texts, and finally, in order to know the frequency of connectors, ***CONCORDANCE > Statistical Analyses*** is applied. The statistical analysis of connectors is displayed in Fig. 2.

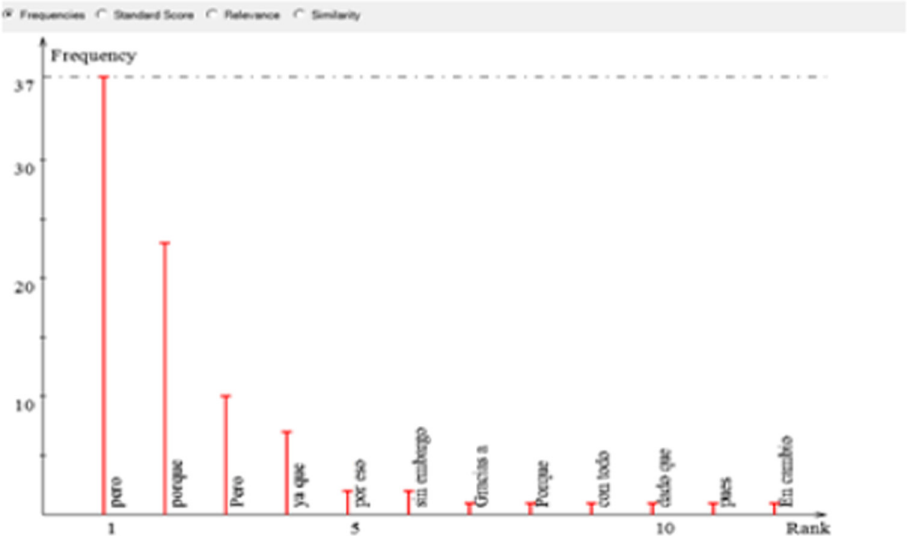

Fig. 2. Frequency of discourse connectors in the corpus

This statistical analysis shows that the counterargumentative discourse connector *pero* 'but' ranks first as the most frequent connector, and in the second place stands the causal discourse connector *porque* 'because'. Thus, our analysis with ***TEXT > Locate*** finds 36 occurrences of causal discourse connectors and 51 occurrences of counterargumentative discourse connectors, showing a clear tendency. And conversely, it does not find any occurrence of consecutive discourse connectors. Our interpretation of this lack of consecutive discourse connectors is related to the fact that orality prevails over writing, as we have already indicated in the previous section.[2]

The low frequency of consecutive discourse connectors, as for example, *ya que* 'since', is quite striking and points out the need to reinforce the use of this type of discourse connectors among students. However, we will not indicate here how to increase their use, and leave this challenge to future work.

3 Our NooJ Grammars. Analysis of Sentences Containing Discourse Connectors

At this point, we are interested in addressing certain phrases containing discourse connectors with our NooJ syntactic grammars. Thus, we will analyse three sentences of our corpus.

- Sentence 1:
  - pero la verdad (es)[3] que no soy de tener amigos prefiero estar solo.
  - 'but the truth (is) that (I) am not inclined to have friends (I) prefer to be alone.'

This sentence shows significant complexity, since it begins with the extrasentential counterargumentative discourse connector *pero* 'but', and then two asyndetic coordinate clauses follow: *la verdad* (es) *que no soy de tener amigos* 'the truth (is) that (I) am not inclined to have friends', and *prefiero estar solo* '(I) prefer to be alone'. The coordinator has been omitted, but it can be easily deduced from the sentence structure or the context. A syntactic grammar is created with NooJ, which includes embedded graphs for the sake of clarity (Fig. 3).[4]

An additional detail is that the first person pronoun subject has been dropped, that is, there is a PRO_DROP null or implicit subject in the subordinate clause of the first

[2] Some occurrences are not, properly speaking, discourse connectors. For example, *con todo el corazón*, "with all the heart", where *todo* is the first modifier of the noun *corazón* inside the noun phrase following the preposition *con*, is an adverbial of manner, whereas in the sentence *no tenía dinero pero con todo pagó el taxi*, "(she/he/it) didn't have money but all the same (she/he/it) paid the taxi," *con todo* certainly is a counterargumentative discourse connector which means "all the same". However, for reasons of space, we will not analyse the differences between these phrases.

[3] The verb has been omitted in the original sentence of the corpus, but we supply the expected third person singular of the verb *ser* 'to be' so as not to make the grammar more complex.

[4] In this particular grammar, we had no difficulty in creating embedded graphs inside embedded graphs up to three levels.

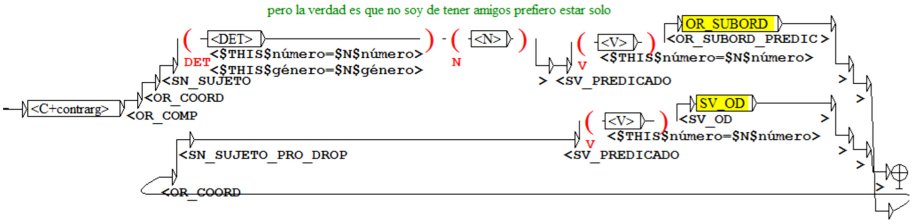

Fig. 3. Main grammar of Sentence 1

coordinate sentence: *que no soy de tener amigos* '(I) am not inclined to have friends' and also in the second coordinate clause *prefiero estar solo* '(I) prefer to be alone'.[5]

The first embedded graph (OR_SUBORD), corresponding to the subordinate clause, which also has an embedded graph inside it, is displayed in Fig. 4. For reasons of space, we do not include here the graph of the verb phrase SV.

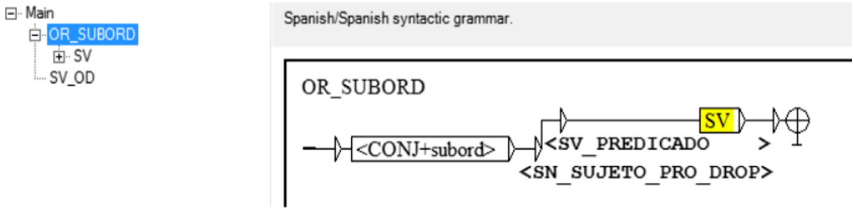

Fig. 4. Graph of OR_SUBORD embedded into the main grammar of Sentence 1

Next, the graph corresponding to the prepositional phrase SP is displayed in Fig. 5 in order to show how it is possible to embed graphs recursively. We do not include here the graph (SV_OD) inside the second coordinate clause, which is the direct object of the main verb *prefiero*, because we have already shown the syntactic richness of the first sentence by displaying the previous figures.

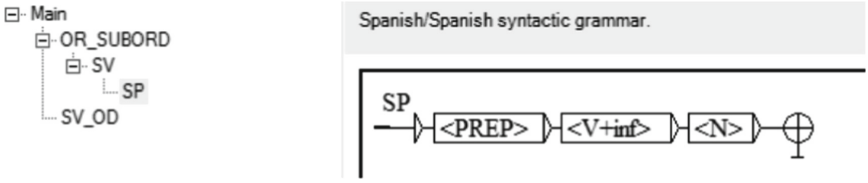

Fig. 5. Graph of SP embedded into the embedded graph SV

[5] Here, we again take up the advancements we achieved in previous work [6] in the processing of sentences with an empty element.

These grammars show how it is possible to adapt NooJ's graph editor to describe syntactic constructions typical of young people's writing, without coordinators, or punctuation marks, but that can likewise be analysed as they were originally produced.

Now let us continue with Sentence 2.

- Sentence 2*:*
  - pues yo si tengo amigas pero también prefiero estar sola.
  - 'well yes I have friends but (I) also prefer to be alone.'

This second sentence begins with the extrasentential causal discourse connector *pues* 'well' and has a predicate with two nucleus verb phrases, *tengo* 'have' and *prefiero* 'prefer', coordinated by the intrasentential counterargumenative discourse connector *pero* 'but'. Its grammar is displayed in Fig. 6.

Fig. 6. Grammar of Sentence 2

After applying ***Show Debug*** in order to check if the grammar works, we perform Generation (Fig. 7) to see the structure of this second sentence.

```
# Dictionary
#
# Language is: sp
#
# Alphabetical order is not required.
#
# Use inflectional & derivational paradigms' description files (.nof), e.g.:
# Special Command: #use paradigms.nof
#
# Special Features: +NW (non-word) +FXC (frozen expression component) +UNAMB (unambiguous lexical entry)
#                   +FLX= (inflectional paradigm) +DRV= (derivational paradigm)
#
# Special Characters: '\' '"' ' ' ',' '+' '-' '#'
#
#
# Dictionary generated automatically
#

<C+caus> <PRON+pers> <ADV> <V> <N> <CONJ+coord> <ADV> <V> <V+inf>


<ADJ>,OR_COMP#<OR_COORD#<SN_SUJETO>#<SV_PREDIC<$THIS$número=$PRON$número><OD>#>#>


<OR_COORD#<SN_SUJETO_PRO_DROP>#<SV_PREDICADO<$THIS$número=$N$número><SV_OD>#>#>#
```

Fig. 7. Performing *Generation* in the grammar of Sentence 2

- Sentence 3:
  - A Pablo le encantaba ir al parque pero había un problema.
  - 'For Paul (it) was enjoyable for him to go to the park but (there) was a problem.'

There is further complexity in this sentence that includes two clauses coordinated by the intrasentential counterargumentative discourse connector *pero* 'but'. The subject of the first coordinate clause is the subordinate clause *ir al parque* 'to go to the park', whereas the second coordinate clause, *había un problema* '(there) was a problem', is impersonal and has no subject at all. The grammar of Sentence 3 is shown in Fig. 8.

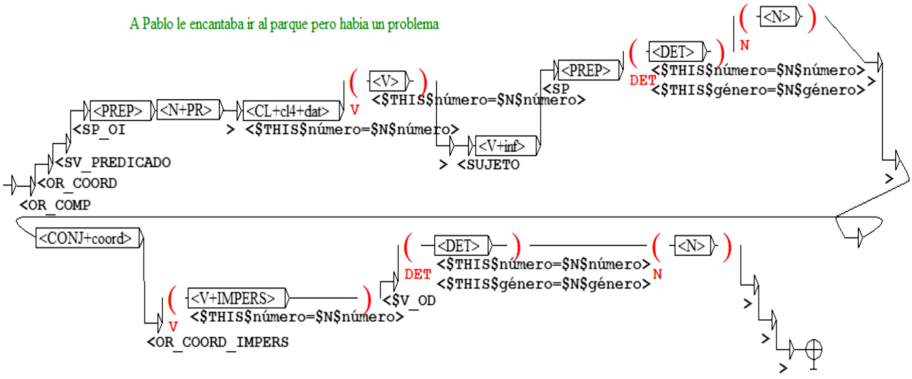

Fig. 8. Grammar of Sentence 3

To check that this grammar suits Sentence 3 we perform ***TEXT > Linguistic Analysis*** (Fig. 9 and Fig. 10).

A Pablo le encantaba ir al parque pero*había un problema.*

A Pablo le encantaba ir al parque ***pero había un problema.***

Finally, as our check is successful, we apply ***Locate*** and then ***Outputs***, so that its structure is shown (Fig. 11).

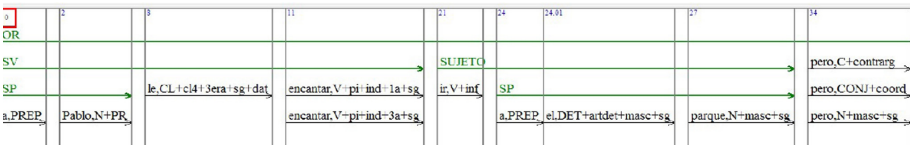

Fig. 9. First clause of Sentence 3

This last figure displays the complexity of the third sentence by highlighting the possibility of including an impersonal construction in a NooJ grammar.

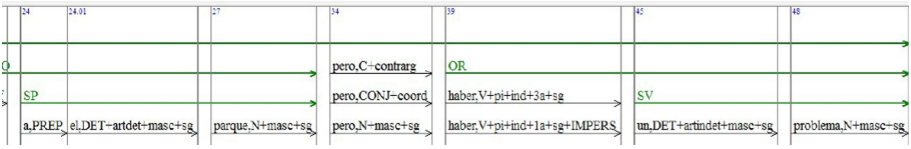

Fig. 10. Second clause of Sentence 3

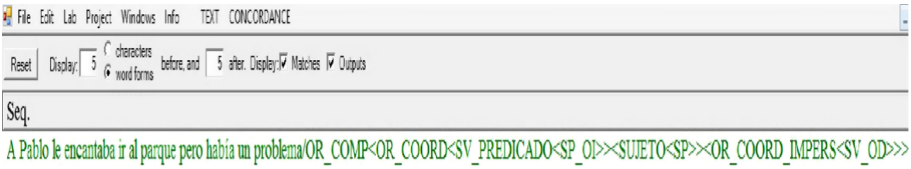

Fig. 11. Structure of Sentence 3

4 Conclusions

This paper is mainly based on Silberztein (2015) and Silberztein (2016). In the first place, as we only had the tag [C+caus] to identify causal discourse connectors, we introduced two new tags in our dictionary, [C+consec] and [C+contrarg], in order to analyse consecutive and counterargumentative connectors, respectively. Corpus analysis showed that counterargumentative connectors greatly outnumber causal connectors and that there is almost a striking absence of consecutive connectors.

In the second place, we analysed certain sentences of the corpus. This analysis gave us an interesting view of the degree of plasticity of NooJ's graph editor, since it allowed us to account for various sentences and clauses: coordinate clauses, subordinate clauses inside coordinate clauses, clauses with null subject, and predicates with more than one nucleus verb phrase. We not only analysed different sentences, but we were also able to apply Generation, among other resources, in order to validate our initial linguistic description. All this work relied on our research project on the pedagogical application of NooJ [5].

Metalinguistic reflection as the generator of learning provides support for our research. In our study, we show how it is possible to process with NooJ the texts produced by tertiary students who will become primary education teachers. We checked how chat or textspeak writing licence (deviating from conventions) stands at the same level of expressions that are close to orality and make syntactic structure extremely rich. The analysis of this corpus leads us to consider our role as teachers in terms of how necessary it is to intervene in the learning process, what aspects may be changed, and what aspects should be maintained and fostered.

References

1. Martín Zorraquino, M.A., Portolés Lázaro, J.: Los marcadores del discurso. In: Bosque, I., Demonte, V. (eds.) Gramática descriptiva de la lengua Española, pp. 4051–4213. Espasa-Calpe, Madrid, España (1999)

2. Montolío, E.: Conectores de la lengua escrita. Ariel Letras. Barcelona, España (2015)
3. Odrowąż-Coates, A.: Socio-Educational Factors and the Soft Power of Language: The Deluge of English in Poland and Portugal. Rowman & Littlefield, Maryland (2019)
4. Ramírez, S.Q.: Identificación de los conectores discursivos de más alta frecuencia en notas periodísticas deportivas. RLA. Revista de lingüística teórica y aplicada, **53**(2), 47–71 (2015). https://scielo.conicyt.cl/pdf/rla/v53n2/art_03.pdf
5. Rodrigo, A., Bonino, R.: Aprendo con NooJ. Editorial Ciudad Gótica, Rosario (2019)
6. Rodrigo, A., Reyes, S., Mota, C., Barreiro, A.: Causal Discourse Connectors in the Teaching of Spanish as a Foreign Language (SFL) for Portuguese Learners Using NooJ. In: Fehri, H., Mesfar, S., Silberztein, M. (eds.) NooJ 2010. CCIS, vol. 1153, pp. 161–172. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-38833-1_14 ISBN 9783030388324
7. Silberztein, M.: La formalization des langues, l´approche de NooJ. ISTE Editions, London (2015)
8. Silberztein, M.: Formalizing Natural Languages: The NooJ Approach. Wiley, London (2016)
9. Spanish Module Argentina. https://www.noojassociation.org/resources.html. Accessed 10 Apr 2020