BRANEN and BRANES Corpora

Amanda Maraschin Bruscato, University of Algarve, Portugal
Jorge Baptista, University of Algarve, Portugal

The European Conference on Language Learning 2021
Official Conference Proceedings

Abstact
This paper presents two learner corpora built to investigate anaphora across learning environments: the *Brazilian Learners of Anaphora in English* (BRANEN) and the *Aprendices Brasileños de Anáfora en Español* (BRANES). Texts were written by language undergraduate students during an online course on anaphora, offered at a Brazilian University in 2020. The corpora provide insights for the analysis of the learning process of anaphora in English and Spanish by Brazilian Portuguese native speakers with intermediate-advanced level in the foreign language. Informants are 30 English and 15 Spanish learners, who were randomly divided into three sub-groups: one group that had two synchronous lessons on anaphora; another that had two asynchronous lessons; and a control group that did not take any lessons. Each participant wrote 100-150 words as a conclusion of a short story. The exercise was performed in four moments: before the course started, after the first lesson, after the second lesson, and a month after the course ended. The texts are available on *Sketch Engine*, a corpus manager and text analysis software, and contain information about the participants' group and testing moment. The BRANEN corpus was automatically part-of-speech tagged with the *Modified English TreeTagger* and has 120 documents, 1,069 sentences, and 1,678 lemmas. For BRANES corpus, the *Spanish FreeLing tagset* was used, and it consists of 60 documents, 543 sentences, and 1,299 lemmas. The *Concordance* tool was used to retrieve sentences with pronominal and zero anaphora, which were then manually and independently annotated by two anaphora experts.

Keywords: Anaphora, BRANEN, BRANES, Learner Corpora, Learning Environments

Introduction

This paper presents two new learner corpora built to investigate the learning of anaphora: the *Brazilian Learners of Anaphora in English* (BRANEN) and the *Aprendices Brasileños de Anáfora en Español* (BRANES). They are available on *Sketch Engine* and are necessary to consider the learning environment as a variable when analysing the learning of anaphora by Brazilian learners of English and Spanish.

Anaphora is an important cohesive mechanism (Halliday & Hasan, 1976), and its knowledge is indispensable for communication in the language. Instead of overusing nominal repetition, as in (1a), speakers can use pronouns (1b), or ellipsis (1c), to refer to the antecedent in the text. While zero anaphora (1c) is commonly used in English in coordinate clauses with the same subject, it is more used in Null Subject Languages (Chomsky, 1981; Rizzi, 1982) as Portuguese and Spanish.

(1a) *Anna*i *wakes up every morning and Anna*i *goes to work.*
(1b) *Anna*i *wakes up every morning and she*i *goes to work.*
(1c) *Anna*i *wakes up every morning and Ø*i *goes to work.*

There are some corpora built specifically to investigate anaphora, such as *OntoNotes* (Pradhan *et al*., 2007), *Anaphora Resolution and Underspecification* (Poesio & Artstein, 2008), *WikiCoref* (Ghaddar & Langlais, 2011), and *Zero Anaphora Corpus* (Baptista, Pereira, & Mamede, 2016). However, they are not learner corpora.

Learner corpora are necessary in Linguistics to give insight on what learners really produce in the target language. Although there are also several of such corpora, as the *International Corpus of Learner English* (Granger, 2003), the *Multilingual Learner Corpus* (Tagnin & Fromm, 2009), and the *Corpus Escrito del Español L2* (Lozano & Mendikoetxea, 2013), they usually do not consider the learning environment in their compilation.

MiLC (Andreu *et al*., 2010) is a multilingual learner corpus that contains written synchronous and asynchronous computer-mediated communication texts. It was used to analyse interlanguage errors in online teleconferences and emails and, though it considers synchronous and asynchronous communication, it does not consider the learning progress in different learning environments.

The BRANEN and BRANES corpora were developed as part of a PhD research aiming at analysing the effectiveness of distance learning modalities to teach anaphora in foreign languages. Thus, texts were collected during a short online course offered to language students. To analyse the effectiveness of the course, participants were randomly divided into three sub-groups: one group that had two synchronous lessons on anaphora; another that had two asynchronous lessons; and a control group that did not take any lessons. The method of data collection and information about participants will be detailed in the next section. Then, the corpora will be described.

Method

Once the learner corpora available usually do not consider learning environments within the criteria guiding their compilation, we have planned a distance course on anaphora to collect

learners' written texts and compile the corpora we needed for our research on the learning of anaphora.

A two-week distance course on the subject was taught in the first semester of 2020 to undergraduate students with a major in English or Spanish at a Brazilian university. After the informed consent was obtained, students answered a grammar questionnaire with 20 questions from Cambridge University or Cervantes Institute assessment questionnaires (equally distributed among levels A2, B1, B2, and C1) to check their proficiency level in the target language. Only those with an intermediate-advanced level attended the course.

A total of 45 learners participated in this longitudinal study. The majority is female (73%) and the median age was 20 years old (they were between 18 and 41 years). Most of them studied English (67%) at the university and were in the third semester (62%) of their bachelor's degree. The others studied Spanish and were in the fifth semester.

For each language, participants were randomly divided into three sub-groups: one group that had two synchronous lessons on anaphora; another that had two asynchronous lessons; and a control group that did not take any lessons. The university e-learning platform (Moodle) was used for the lessons. Each synchronous lesson used videoconference for 90 minutes; for asynchronous lessons, short videos, texts, discussion forums, and automatic exercises were used. The course design is further detailed in Bruscato and Baptista (2021).

The proficiency test results are presented in Table 1.

Language	Group	Mean	Standard deviation
Spanish	Asynchronous (N=5)	15	2
	Synchronous (N=5)	15	2
	Control (N=5)	14	5
English	Asynchronous (N=10)	14	3
	Synchronous (N=10)	12	4
	Control (N=10)	13	5

Table 1: Proficiency Test Results

Students wrote narratives with 100-150 words to conclude short stories and submitted their text files on Moodle. The task was planned to control the possible antecedents in the texts. To track learning over time, the exercise was performed in four moments: before the course started, after the first lesson, after the second lesson, and a month after the course ended. Table 2 presents the four stories' human antecedents.

	English	Spanish
Test 1	John, Mary, twins Joseph, Ana, parents Witch, neighbour	Juan, María, gemelos José, Ana, padres Bruja, vecina
Test 2	Luke, Louis, brothers Liz, mother	Lucas, Luís, hermanos Liz, madre

	Ariel, person	Ariel, persona
Test 3	Matthew, Laura, couple Manuel, Leonard, fathers Father Augusto	Mateus, Laura, pareja Manuel, Leonardo, padres Fray Augusto
Test 4	Alice, Helena, friends, neighbours Other girls Anthony, classmate	Alice, Helena, amigas, vecinas Otras chicas Antonio, compañero de clase

Table 2: Human Antecedents

The corpora are available on Sketch Engine[1] (Kilgarriff *et al*., 2004), a corpus managing and text analysis software, and they include metadata about the participants' group (asynchronous, synchronous, control) and testing moment (1, 2, 3, 4). The Sketch Engine corpus query system was chosen because it is used by linguists, teachers, and students all over the world and also because the European Lexicographic Infrastructure project will provide all academic institutions in the EU free access to the software, at least until 2022.

The BRANEN corpus has 120 documents and was automatically part-of-speech (POS) tagged with the Modified English TreeTagger, while the BRANES corpus has 60 documents and was POS-tagged with the Spanish FreeLing tagset. In the next section, some results from the corpora will be presented.

Corpora

We have mainly used two Sketch Engine tools to analyse the corpora: Wordlist and Concordance. Wordlist was used to reveal the most frequent nouns in the corpora, which, as expected, are the antecedents presented in the first part of each story. We have checked the Spearman's rank correlation coefficient to compare the nouns used by the asynchronous (Asyn), synchronous (Syn), and control (Cont) groups. As Table 3 shows, there was a moderate-to-high correlation.

	English			Spanish		
	Asyn-Syn	Asyn-Cont	Syn-Cont	Asyn-Syn	Asyn-Cont	Syn-Cont
Test 1	0.619	0.684	0.721	0.787	0.807	0.743
Test 2	0.663	0.749	0.734	0.717	0.872	0.710
Test 3	0.630	0.682	0.686	0.739	0.780	0.666
Test 4	0.656	0.681	0.674	0.832	0.831	0.704
All nouns	0.720	0.772	0.747	0.771	0.780	0.689

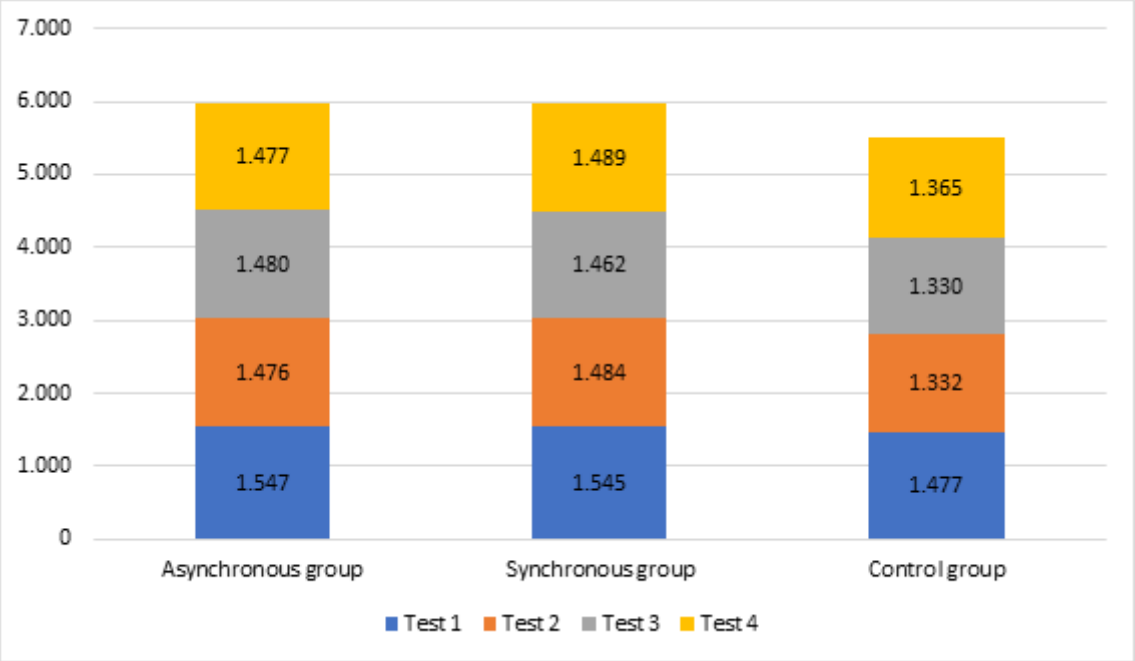
Table 3: Spearman-coefficient

BRANEN

The BRANEN corpus contains texts written by 30 English learners in 4 moments of their learning process. In total, it has 1,069 sentences, 1,678 lemmas, 2,242 unique words, 17,454 words, and 19,934 tokens. Graph 1 shows the breakdown of the total number of words by each group -- asynchronous (Asyn), synchronous (Syn), control (Ctrl) -- and in each test (1, 2, 3, 4). The relative size of each group across tests (1 to 4) is highly correlated: Pearson(Asyn,Syn)=0.968, (Asyn,Ctrl)=0.927, and (Syn,Ctrl)=0.970. The same holds for the

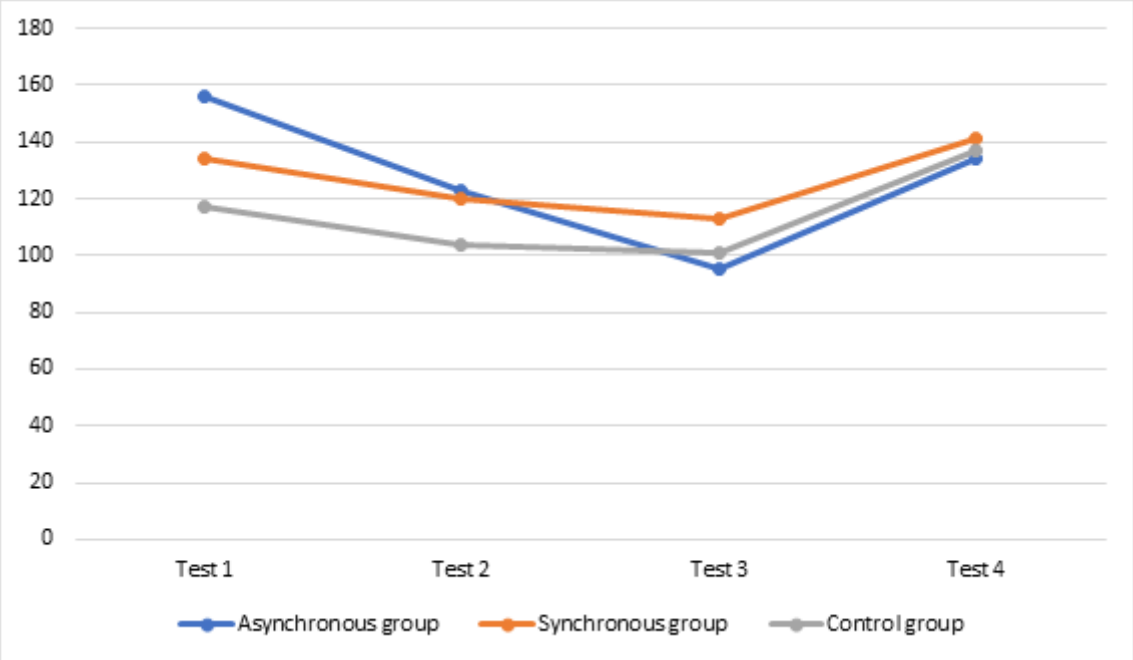
[1] www.sketchengine.eu

relative size of each test across groups, ranging from Pearson (T3-T4)=0.980 to T2-T4)=0.999. Although the control group wrote slightly less words than the other groups, the difference was not significant. This was already expected, since participants were told to conclude each story with between 100 and 150 words.



Graph 1: BRANEN Word Distribution

We then checked if there was a difference in the use of pronominal anaphora among groups over time. Thus, we have checked the frequency of subject pronouns in the corpus, as they can be distinguished for their case. The results are presented in Graph 2. The synchronous and control groups have similar results, with a slight reduction in the use of subject pronouns until Test 3 and an increase in Test 4, which was completed one month after the end of the course. For the asynchronous group, there was a significant decrease until Test 3, but again an increase in Test 4. It is possible that the asynchronous group learned how to replace the pronouns with ellipsis or nouns, throughout the course, as it will be shown below. This trend, however, was not kept, as the data from Test 4 shows.



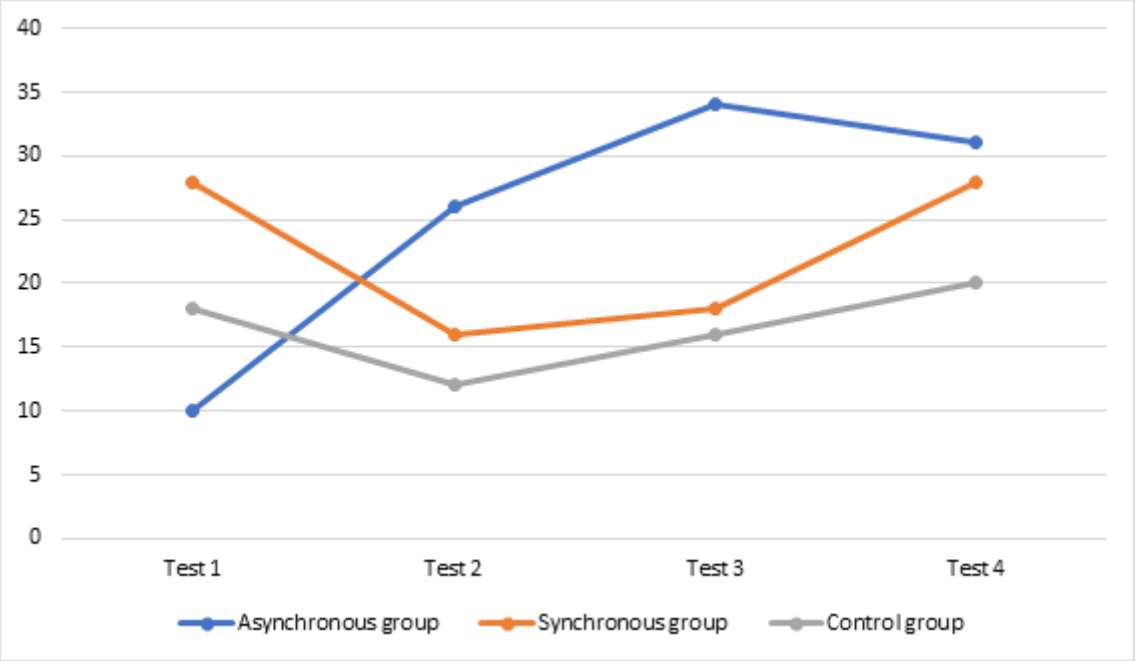
Graph 2: Subject Pronouns Frequency

An initial study was also carried out with the Sketch Engine tools, to compare the use of different types of anaphora over time, using the Concordance Corpus Query Language (CQL). The frequency of nominal, pronominal, and zero anaphora in coordinate clauses was considered. Coordinate clauses were chosen because they allow ellipsis in the subject position, a phenomenon that has been explicitly addressed during the course. At this time, only sentences with a coordinate conjunction [tag="CC"] were retrieved, when followed either by a noun [tag="N.*"] or a personal pronoun [tag="PP"] and then a verb [tag="V.*"]. Ellipsis was captured when the verb immediately follows the coordinate conjunction. Results are presented in Table 4 and, as Graph 3 highlights, there was an increase in the use of zero anaphora by the asynchronous group.

Group	Test	Nominal anaphora	Pronominal anaphora	Zero anaphora
Asynchronous	1	4	6	10
	2	8	6	26
	3	8	3	34
	4	10	10	31
Synchronous	1	7	6	28
	2	9	5	16
	3	11	7	18
	4	13	10	28
Control	1	8	4	18
	2	9	4	12
	3	13	8	16

	4	11	5	20
Total		111	74	257

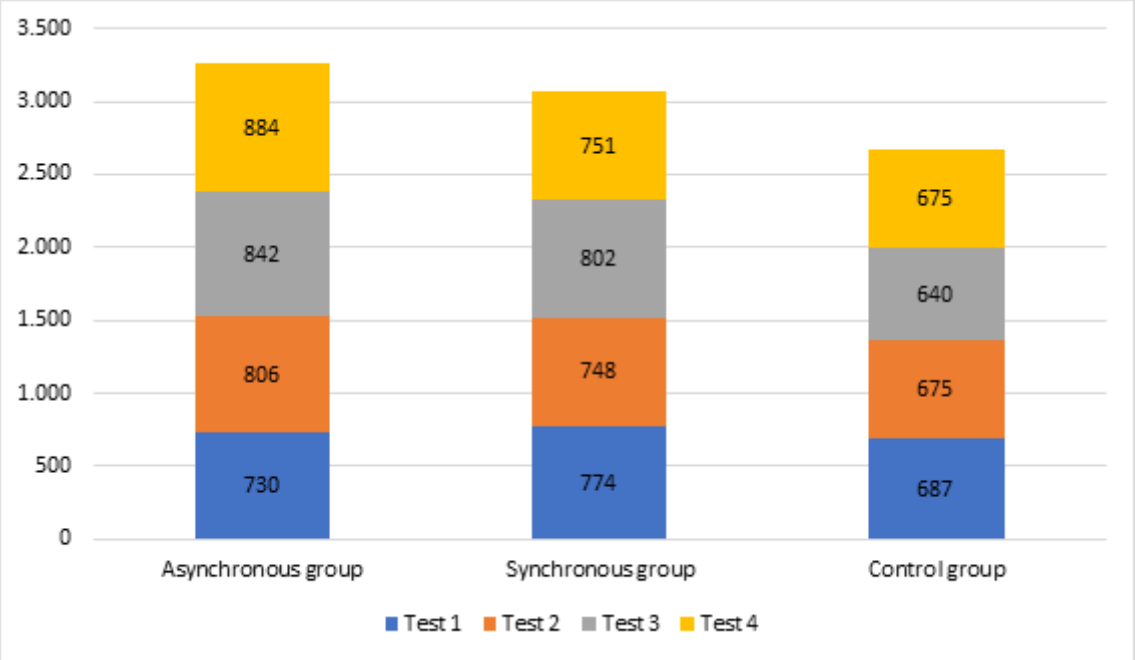
Table 4: Subject Anaphora Types in English Coordinate Clauses



Graph 3: Zero-Anaphora Subjects in Coordinate Clauses

BRANES

The BRANES corpus contains texts written by 15 Spanish learners in 4 moments. In total, it has 543 sentences, 1,299 lemmas, 2,095 unique words, 9,021 words, and 10,233 tokens. Graph 4 shows the word distribution for each group in each test. The number of written words constantly increased for the asynchronous group. While the asynchronous group wrote more than the others, the control group wrote less.



Graph 4: BRANES Word Distribution

As with BRANEN, Wordlist and Concordance were used to check the frequency of nominal, pronominal, and zero anaphora. When we searched specifically for subject pronouns, only 126 occurrences were found in the whole corpus. Due to the small number, their use over time was not analysed here. When we look directly to subject anaphora in coordinate clauses, as Table 5 shows, it is clear that learners predominantly use zero anaphora. In this matter, they behave like native speakers, probably because Portuguese and Spanish are both Null Subject Languages.

Group	Test	Nominal anaphora	Pronominal anaphora	Zero anaphora
Asynchronous	1	0	0	7
	2	2	0	8
	3	1	0	5
	4	2	0	7
Synchronous	1	3	2	10
	2	1	0	9
	3	1	0	8
	4	0	2	11
Control	1	1	0	9
	2	1	0	5
	3	3	0	9
	4	4	0	7
Total		19	4	95

Table 5: Subject Anaphora Types in Spanish Coordinate Clauses

Ongoing Analysis

After these preliminary results, the human subjects in the corpora were annotated using Recogito annotation tool[2]. This software was chosen because it allows us not only to annotate words and phrases, but also to establish unilateral (oriented) relations between them. Besides, it is freely available online and it is possible to involve multiple annotators in the task. One can also export the results into csv files and make the corpora available to the public.

Figures 1 and 2 show narratives written by English and Spanish learners in the first test. The anaphors and their antecedents are highlighted in different colours according to their category (proper nouns, common nouns, pronouns, or ellipsis). To annotate zero anaphora, we have marked the corresponding verbs (which are also highlighted). The anaphoric relations between the anaphors and their antecedents are established as intra (*in*) or trans (*tr*) sentential. In the examples below, we can see that, while the English learner does not use null subjects, the Spanish student does not use subject pronouns.

[2] https://recogito.pelagios.org/

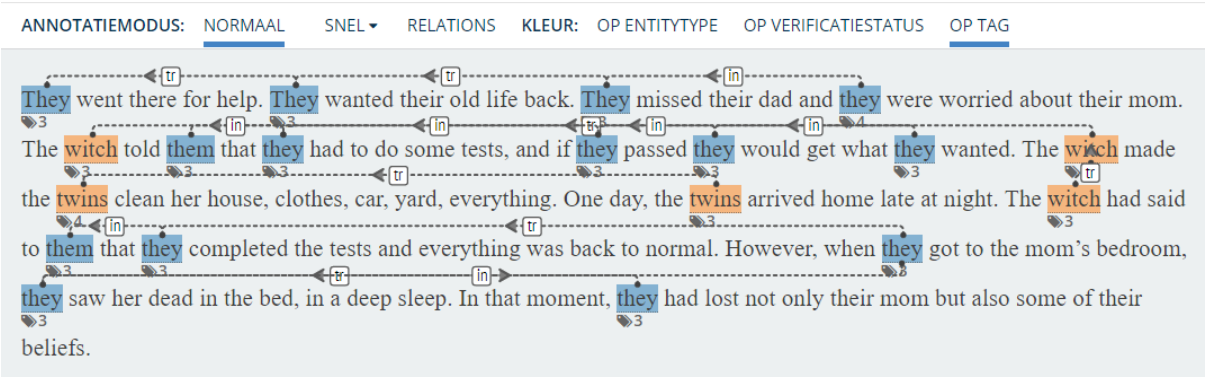


Figure 1: Subject Anaphora Annotation in English Narrative

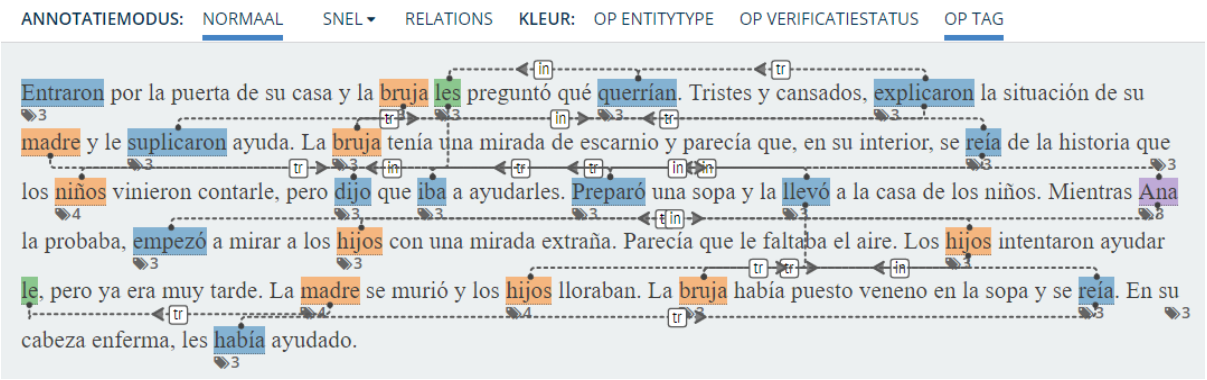


Figure 2: Subject Anaphora Annotation in Spanish Narrative

To check the consistency of the annotation, part of the corpora received a second, independent annotation, following a set of comprehensive guidelines, and then were compared for assessing inter-rater agreement. The analysis of the annotation will be discussed in a future paper, where we will also compare learners' texts with natives' texts for the same task. After the publication of the results, we will make the annotation available to the public.

Conclusions

The aim of this paper was to present BRANEN and BRANES, two new longitudinal learner corpora built to investigate the impact of the distance synchronous and asynchronous learning of anaphora in English and Spanish. While describing the corpora, we were also able to share some preliminary results.

Since Portuguese and Spanish are both null-subject languages, learners already used ellipsis since the first test. In English, however, those who participated in the asynchronous course learned that it is better to use ellipsis in coordinate clauses with the same subject than to repeat the noun or pronoun.

The asynchronous group performed better probably because they had to write more during the course, while participating in written discussion forums. On the other hand, the synchronous group participated orally during videoconferences. It is possible that, if we had included spoken data, evidence could have been gathered showing that the synchronous group had performed better in this respect.

In this study, we have focused on written data and have analysed only subject anaphors. A future step would be to include spoken data from the video recordings and to investigate other anaphors. Hopefully, this first version of the corpora will help other researchers, teachers, and learners to better investigate and understand the learning process of anaphora in foreign languages.

Similar initiatives can also be undertaken focusing on other linguistic phenomena pertinent to L2 learning, for example, adverb placement (Rankin, 2010).

BRANEN and BRANES are available on Sketch Engine and contain metadata about the participants' group and testing moment. After we conclude the annotation analysis, the corpora will also become available to the public on Recogito.

The results here presented, though preliminary, show that considering the learning environments in learner corpora studies is certainly a relevant variable. To the best of our knowledge, BRANEN and BRANES are the first corpora to take this factor into account in their design.

Acknowledgements

We would like to thank the University of Algarve for the funding to participate in this conference, as well as the professors and students of the Language Institute of the Federal University of Rio Grande do Sul for their collaboration in this research, in particular Dr. Sergio Menuzzi, the Director of the Faculty, for the support in the implementation and accreditation of the course.

References

Andreu, M., Astor, A., Boquera, M., MacDonald, P., Montero, B., & Pérez, C. (2010). Analysing EFL learner output in the MiLC project: An error it's*, but which tag. *Corpus-based approaches to English language teaching*. London: Continuum, 167-179.

Baptista, J., Pereira, S., & Mamede, N. (2016) ZAC: Zero Anaphora Corpus. *In* Fontes, H.L. & Branco, A. (eds.) Workshop on Corpora and Tools for Processing Corpora (co-located with PROPOR 2016), 38-45. Retrieved from: https://ec.europa.eu/translation/portuguese/magazine/documents/folha50_propor_papers_en.pdf

Bruscato, A. M., & Baptista, J. (2021). Designing an Online Course to Teach Anaphora in Foreign Languages. *Proceedings of The 3rd International Conference on Teaching, Learning and Education*, 57-68. Retrieved from: https://www.dpublication.com/proceeding/3rd-ictle

Bruscato, A. M., & Baptista, J. (2021). Synchronous and Asynchronous Distance Learning of Anaphora in Foreign Languages. *Texto Livre, 14*(1), 1-16. Retrieved from: https://periodicos.ufmg.br/index.php/textolivre/article/view/29177/26443

Chomsky, N. (1981) *Lectures on Government and Binding: The Pisa Lectures*. Dordrecht: Foris.

Ghaddar, A., & Langlais, P. (2016). Wikicoref: An English coreference-annotated corpus of Wikipedia articles. In *LREC'16*, 136-142.

Granger, S., Dagneaux, E., and Meunier, F. (2002). *International Corpus of Learner English*. Louvain: UCL.

Halliday, M., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). Itri-04-08 the Sketch Engine. *Information Technology,* 105-116.

Lozano, C. (2009). CEDEL2: Corpus Escrito del Español L2. In: Callejas, C. M. B. (Ed.). (2009). *Applied Linguistics Now: Understanding Language and Mind*. Universidad de Almería, 197-212.

Poesio, M., & Artstein, R. (2008). Anaphoric Annotation in the ARRAU Corpus. In *LREC*, 1170-1174.

Pradhan, S. S., Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., & Weischedel, R. (2007). Ontonotes: A unified relational semantic representation. In *ICSC*, 517-526.

Rankin, T. (2010). Advanced learner corpus data and grammar teaching: Adverb placement. *Corpus-Based Approaches to English Language Teaching*. London: Continuum, 205-215.

Rizzi, L. (1982) *Issues in Italian Syntax*. Dordrecht: Foris.

Tagnin, S. E. (2006). A multilingual learner corpus in Brazil. In *Corpus linguistics around the world*, Brill Rodopi, 195-202.

Contact email: amandabruscato@gmail.com