Manuscript
Click here to access/download;Manuscript;Revised_LREV-D-20-00079.tex
Click here to view linked References

Noname manuscript No.
(will be inserted by the editor)

A sequence labelling approach for automatic analysis of *ello*

Tagging pronouns, antecedents, and connective phrases

Giovanni Parodi · Richard Evans* · Le An Ha · Ruslan Mitkov · Cristóbal Jesus Julio Vergara · Raúl Ignacio Olivares López ·

Received: date / Accepted: date

Abstract Encapsulators are linguistic units which establish coherent referential connections to the preceding discourse in a text. In this paper, we address the challenge of automatically analysing the pronominal encapsulator *ello* in Spanish text. Our method identifies, for each occurrence, the antecedent of the pronoun (including its grammatical type), the connective phrase which

Giovanni Parodi
Pontificia Universidad Catolica de Valparaiso, Chile
Tel.: +56 32 227 3000
E-mail: giovanni.parodi@pucv.cl

Richard Evans*
Research Institute of Information and Language Processing, University of Wolverhampton, Wulfruna Street, Wolverhamton, West Midlands, WV1 1LY. United Kingdom.
Tel.: +44 190 232 2924
E-mail: r.j.evans@wlv.ac.uk

Le An Ha
Research Institute of Information and Language Processing, University of Wolverhampton, United Kingdom.
Tel.: +44 190 232 8705
E-mail: l.a.ha@wlv.ac.uk

Ruslan Mitkov
Research Institute of Information and Language Processing, University of Wolverhampton, United Kingdom.
Tel.: +44 190 232 2471
E-mail: r.mitkov@wlv.ac.uk

Cristóbal Jesus Julio Vergara
Pontificia Universidad Catolica de Valparaiso, Chile
Tel.: +56 32 227 3000
E-mail: cristobal.julio@pucv.cl

Raúl Ignacio Olivares López
Pontificia Universidad Catolica de Valparaiso, Chile
Tel.: +56 32 227 3000
E-mail: raul.olivares.lopez@gmail.com

combines with the pronoun to express a discourse relation linking the antecedent text segment to the following text segment, and the type of semantic relation expressed by the complex discourse marker formed by the connective phrase and pronoun. We describe our annotation of a corpus to inform the development of our method and to finetune an automatic analyser based on bidirectional encoder representation transformers (BERT). On testing our method, we find that it performs with greater accuracy than three baselines (0.76 for the resolution task), and sets a promising benchmark for the automatic annotation of occurrences of the pronoun *ello*, their antecedents, and the semantic relations between the two text segments linked by the connective in combination with the pronoun.

Keywords anaphora resolution · encapsulation · ello · referential coherence · relational coherence

1 Introduction

In written discourse comprehension, one of the most intricate textual relations from a psycholinguistic perspective is the anaphoric resolution of neuter pronouns, which are linguistic particles devoid of lexical meaning and which lack morphosyntactic features that may guide the reader. Particularly crucial to the comprehender are discourse constructions in which s/he must not only establish a coherent referential connection to the preceding discourse, but also and conjointly, infer and identify a semantic relation between the essential discourse segments of, for example, causal semantic relations.

Readers invest great cognitive effort in comprehending a written text and constructing a coherent mental representation of the events described therein [12,24,34]. In Spanish, the neuter pronoun *ello* tends to occur regularly in multi-layered constructions in which it operates in concurrence with other linguistic particles in order to construct and convey complex meanings. Disambiguating the antecedent of a neuter pronoun is essential to progress in the text that is being read. Without properly identifying the referential link as well as the semantic rhetorical discourse relation, it would be difficult for readers to build a complete and coherent representation - those derived representations would fail to capture the genuine meaning of the text.

Referential inferences guided by gendered and numbered features are automatic (as in the case of pronouns such as *he* or *she*) [22,23]. Neuter pronouns such as *ello* that encapsulate one or two clauses and that refer to abstract entities in complex nominalisations require longer reading times and exert reinspections of the possible target antecedents, exploiting more strategic and delayed processes [37]. The psycholinguistic requirements imposed by a neuter pronoun *ello* when it refers to a long antecedent and, at the same time, combines with a connective phrase to form a complex discourse marker expressing a counter-argumentative semantic relation are more demanding than when the pronoun is linked to a short antecedent and combines with a connective phrase expressing a causal relation [38].

This is the first study presenting an automatic method to identify, for each occurrence of the pronoun *ello* in naturally occurring Spanish texts, the antecedent of the pronoun, the syntactic type of antecedent, the connective phrase which is combined with the pronoun to form a complex discourse marker linking the antecedent to the subsequent text segment, and the semantic relation expressed by the discourse marker. In previous research, few studies have focused on description of the pronoun *ello* in a specialised corpus of academic/pedagogic Spanish and even fewer have explored the comprehension processes involved in the different contextual constructions in which it may occur.

The paper is organized as follows. In Section 1.1, we discuss the linguistic phenomenon of encapsulation exhibited by the pronoun *ello*. In Section 2, we describe our approach to the development of an automatic method to analyse and process examples of the pronoun *ello* occurring in input texts. This processing includes determination of the antecedent of the pronoun, the multi-token portion of text that it encapsulates, the linguistic type of its antecedent, identification of the connective phrase which combines with the pronoun to form a complex discourse marker expressing a semantic relation between the antecedent and the subsequent text segment, and classification of each occurrence of *ello* with respect to the type of semantic relation expressed by the discourse marker. In Section 3.1, we describe a corpus that we annotated with information about occurrences of the pronoun that occur within it, the antecedent of each pronoun, and the semantic relations between the antecedent and subsequent text segments. This includes details about the annotation scheme (Section 2.1.1) and formatting of the corpus for use with machine learning methods to perform sequence tagging. In Section 2.3, we frame the anaphora resolution process as a sequence labelling task and present a new neural method exploiting BERT language representations to perform it. In Section 3, we present the results of our corpus development process and the accuracy of the practical systems we developed to automatically classify and identify the antecedents of the pronoun *ello*.

1.1 The neuter pronoun *ello* as encapsulator: Referential and relational coherence

Encapsulation is a text mechanism of cohesion and coherence through which the meaning of textual segments is condensed or labelled, establishing a process of reference and substitution by another textual element, such as pronouns and nouns [18,1,2,15]. Encapsulators contribute to textual thematic progression and to referential maintenance by connecting pieces of discourse and helping to construct the coherence of the text [37,38]. In this way, they are fundamental links that also guide comprehension by converting the encapsulated information into shared knowledge available to the reader [9,40, 51,47]. Encapsulated text segments may be presented in preceding or subse-

quent textual units; consequently, an encapsulator may function cohesively as an anaphor or a cataphor [18,50,53].

There is no consensus on the exact types of unit that can be grouped together into the category of encapsulators [1,2,7,9,14,17,18,28,31,32,40,47, 50,51]. In Spanish, encapsulation is a mechanism executed by a variety of linguistic forms which cannot be categorised as a class of words *per se*. This means that the encapsulating role or function of a given word or noun phrase is dependent upon the context [7,15,28,40,50]. There are many denominations and categories through which the different types of encapsulating mechanisms are described, however there is a general consensus that neuter pronouns such as *ello* make up one of these groups. The difficulty in identifying encapsulators lies in the fact that they do not belong to a specific word class in themselves but they do fulfil a particular textual function [7,28,31,32].

In Spanish, the neuter pronoun *ello* has specific features compared with other personal pronouns, mainly in terms of semantics. Although, morphologically, it corresponds to the third person singular, in fact it has no notion of person as it does not refer to any of the participants in the communicative exchange. The pronoun *ello* contains the inherent properties of what Benveniste [6] refers to as the non-person. Consequently, *ello* does not have the same referential nature that can be identified in other gendered and numbered personal pronouns or even in other neuter pronouns. Its condition as a neuter pronoun makes it "an example of a grammatical class of words that express certain abstract notions" [42, p. 24]. In other words, *ello* shares with all of the neuter pronouns the capacity to reproduce groups of "two or more nouns referring to things (not persons)" [5, p. 80]. Because of its lack of conceptual meaning by comparison to other anaphoric resources, the neuter pronoun has greater interpretative dependence, as it refers to "what has just been said" [59, p. 59] in the clause or clauses that precede it. According to [35,36] and to [41] and [13], the pronoun *ello* can be preceded by sentences or neutral nominal groups, as well as by groups of several related non-personal nouns. Besides, *ello* can be preceded by "abstract, often deverbal names interpreted as events or referring to situations or states of things which would more commonly be represented in sentences" [42, p. 303].

The relevance of the neuter pronoun *ello* as a retrospective encapsulating mechanism is very significant in a multidimensional perspective. Besides operating on the linguistic plane as a cohesive device providing texture to discourse, it also executes a vital function on the cognitive dimension [2,3,7,14, 28]. It provides a procedural meaning [9,10,39,40] that restricts, although to a lesser extent than in nominal anaphora, the possible interpretations of the text segments in which *ello* exerts its connective function.

Parodi and Burdiles [35,36] have identified, based on corpus studies of Economics discourse in Spanish, that neuter pronouns tend to encapsulate - mainly in an anaphoric orientation- extensive text segments as antecedents. Example (1) demonstrates part of the problem we are interested in. In this example, different highlighting colours are used to mark the CLAUSE_COMPLEX

antecedent, the COUNTER-ARGUMENTATIVE semantic relation, and the encapsulator:

(1) "La enorme diversidad que existe de libros y revistas, ropa, alimentos y bebidas, son ejemplos de tales ganancias. Es más difícil percatarse de ellas al adquirir medicamentos con marcas registradas que tienen una composición química idéntica a la de otras alternativas genéricas. A pesar de **ello**, algunas personas sí están dispuestas a pagar más por la alternativa de la marca registrada." (Document PUCV-UCSC-2013-041)

["The enormous diversity of books and magazines, clothing, food and beverages are examples of such gains. It is more difficult to see them when you buy brand-name drugs that have the same chemical composition as other generic alternatives. Despite this, some people are willing to pay more for the trademark alternative."]

In (1) there is a complex antecedent of *ello*, which has several embedded clauses (in the original Spanish version, there are four main verbs). In this passage, taken from a text from a specialised corpus of academic/pedagogic Spanish, the reader faces the challenge of connecting the neuter pronoun *ello* to a long and complex antecedent which precedes it. Besides the required anaphoric referential resolution, the reader must also infer the counter-argumentative semantic relation between these two text segments, signalled in this example by the phrasal connective Spanish marker *a pesar de* (despite). Consequently, there is - at the same time - a double or binary procedural marked instruction to the reader, one of referential rank (*ello*) and another of relational status (*a pesar de*). This is the context in which *ello* tends to occur more frequently in Spanish. To summarise, "... what makes the link is not grammar but the addressee's previous assumption about coherence." [40, p. 296].

As can be seen, *ello* as an encapsulator may operate in a binary construction of referential and relational coherence [45,46]. In this study, we are interested in identifying both kinds of discourse relations, paying attention - at the same time - to the kind of antecedent of the neuter pronoun *ello*, and the semantic relation that connects both discourse segments: the preceding textual portion that acts as the encapsulated antecedent and the subsequent discourse unit that functions as the continuation of the semantic relation (see Figure 1). Particularly critical to this binary functional construction are the multipart discourse markers in which *ello* operates and to which contribute the encapsulation function. Figure 1 shows how these two interconnected functions work together, linked by a complex discourse marker.

As can be seen, two semantic relations are exemplified: counter-argumentative (e.g. *a pesar de*; in spite of) and additive (e.g. *además de*; in addition to). Furthermore, the process of encapsulation and anaphoric referential coherence is depicted and the possible varying encapsulated extensions of the previous discourse segment(s) are specified.

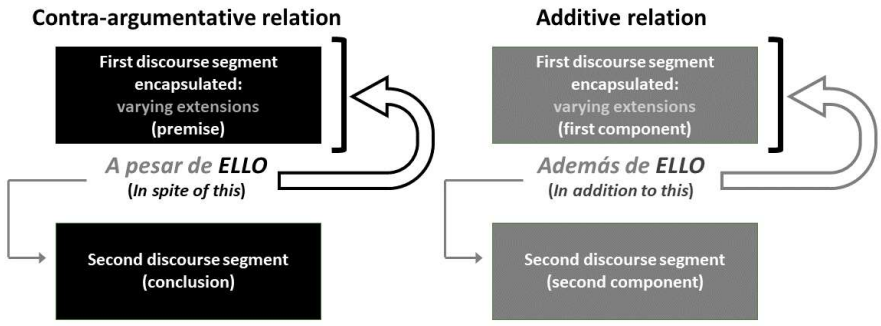

Fig. 1 Interconnected functions linked by a complex discourse marker

In this context, when analysing a complex discourse organisation entrenched in a two-type coherence relation (based on the neuter pronoun *ello*), at least three different features may be identified:

1. the kind of encapsulated anaphoric antecedent (e.g., phrase, clause, clause complex, etc.)
2. the kind of semantic relation involved (e.g., causal, additive, counter-argumentative, etc.)
3. the kind of discourse marker that links the two basic discourse segments and that signals the semantic relation (e.g., *por*, *a pesar de*, *además de*, etc.)

2 Method

Our method to resolve occurrences of the pronoun *ello* to their antecedents is based on three main steps:

1. the development of a corpus in which occurrences of the pronoun *ello*, their antecedents, and the connective phrases which combine with the pronouns to form complex discourse markers linking the preceding and following text segments are annotated with several different types of information,
2. derivation from the annotated corpus of labelled token sequences to support the development of an automatic sequence labelling model,
3. the development of an automatic sequence labelling model to classify occurrences of the pronoun *ello* in input texts and to identify the antecedents of these pronouns.

These are detailed in Sections 2.1–2.3.

2.1 Development of an Annotated Corpus

We compiled a corpus of forty-four handbooks of Economics. They were collected from the reading materials included in the syllabi of two undergraduate

university programmes in Economics and Business Administration in Chile. The texts were digitised in order to be analysed and all figures and graphs were excluded from the analysis. The handbook genre was selected because it represents a specialised discourse that helps students learn theoretical and methodological topics in economics.

We selected 38 documents with a total of 8 243 870 tokens and 350 922 automatically identified sentences. For documents in this corpus, the average number of words was 216 944 (405 331 words for the largest document, 8 968 for the shortest, $\sigma = 111\,500.64$). This corpus was annotated by four trained annotators in the Pontificia Universidad Católica de Valparaíso, who were native speakers of Spanish. They each annotated an equal portion of the data using the PALinkA corpus annotation tool [33] in accordance with the scheme presented in Section 2.1.1. After annotation of the corpus, inter-annotator agreement was assessed as described in Section 3.1.1.

We annotated 2347 of the 2359 occurrences of the encapsulator *ello* occurring in the corpus. Filtering of duplicates and inconsistencies due to tokenisation left 1916 occurrences available for use in the development and evaluation of our sequence labelling approach. These statistics are consistent with the observation that *ello* is rarely used in many text genres (novels, newspapers, letters, etc). However, occurrences of this encapsulator are more frequent in specialised discourse and their functions become more relevant to the comprehension process in text of this type. We believe that procedures for automatic identification of the referent in this complex type of encapsulation can contribute to furthering our understanding of other types of encapsulators. Further, evidence obtained using the Google ngrams viewer indicates that use of *ello* has been increasing over time (from 0.018% in 1800 to 0.030% in 2018). The occurrences annotated in our corpus are those in which *ello* functions as an encapsulator, representing a simple or complex idea, rather than merely substituting a name or concept. These encapsulators were annotated, regardless of whether or not they occur within a connective phrase to explicitly signal a semantic relation.

Our annotation was made by four annotators (one of whom was the fifth author) who were trained by the first author before starting. They were provided with the theoretical framework in which the encapsulator *ello* was analysed and then introduced to the different types of antecedent and the possible semantic relations that may hold between the antecedent and subsequent text segments. This was followed by several joint annotation sessions which led to the derivation of in-house guidelines for the rest of the annotation process. The annotators consulted the guidelines throughout the process.

2.1.1 Annotation Scheme

The annotation scheme was developed on the basis of previous corpus-based studies of the encapsulator *ello* [37]. It includes three markable elements and two relations which may hold between these elements.

The annotation task has two aspects:

1. Tagging of markables as non-empty XML elements, comprising:
   (a) ELLO: An occurrence of the encapsulator *ello* in the corpus. They contain attributes:
      i. ID, specifying an identity number for the tagged pronoun,
      ii. SEMANTIC_RELATION, specifying the semantic relation holding between the preceding and following text segments which is expressed by the complex discourse marker containing the pronoun and the CONNECTIVE_PHRASE.[1] Possible values for this attribute are ADDITIVE, ADVERSATIVE, and CAUSAL. Five different semantic relations were identified in previous work [35], but we only annotated the three listed here which are fundamental for creating coherence and cohesion.
      iii. LINK_ANTECEDENT, indicating the specific ANTECEDENT that the parent ELLO encapsulates. These elements have attributes:
         A. ID: The unique identity number of the relation element.
         B. SRC: The identifying reference number of the ANTECEDENT element.
      iv. TYPE_REF_EXTENT, specifying the type of linguistic unit of its antecedent. Possible values for this attribute are:
         A. NP, denoting noun phrases and including phrases whose heads are non-finite verbs acting as nouns,
         B. CLAUSE, denoting simple sentences with one finite verb,
         C. CLAUSE_COMPLEX, denoting sentences containing two or three clauses linked in coordination or subordination,
         D. TEXT_PORTION, denoting sentences containing four or more clauses linked in coordination or subordination.

      These elements also contain LINK_ANTECEDENT elements/relations (See 2.(a)).
   (b) ANTECEDENT: The portion of text encapsulated by a given instance of *ello*. They contain a single attribute, specifying a unique identity number for the element.
   (c) CONNECTIVE_PHRASE: The portion of text which, in combination with the pronoun, forms a complex discourse marker expressing a semantic relation between the ANTECEDENT and the subsequent discourse segment. They contain a single attribute, specifying a unique identity number for the element. These elements also contain LINK_CONNECTIVE_PHRASE elements/relations (See 2.(b)).
2. Tagging of relations between elements as empty self-closing XML elements. These comprise:
   (a) LINK_ANTECEDENT, indicating the specific ANTECEDENT that the parent ELLO encapsulates. These elements have attributes:
      i. ID: The unique identity number of the relation element.

[1] The complex discourse marker is not explicitly annotated. Only the component pronouns and connective phrases are annotated.

ii. SRC: The identifying reference number of the ANTECEDENT element.

(b) LINK_CONNECTIVE_PHRASE, indicating the specific ELLO that the parent CONNECTIVE_PHRASE links to an ANTECEDENT. These elements have attributes:

i. ID: The unique identity number of the relation element.
ii. SRC: The identifying reference number of the ELLO element.

This scheme was applied to the corpus described in Section 2.1.

2.2 Derivation of Labelled Token Sequences to Build a Sequence Labelling Model

We used *spaCy*[2] [20] to automatically mark up additional linguistic information in our corpus. After processing using spaCy, the corpus annotation encoded information on the tokens in the corpus, their parts of speech, their lemmas, and our manually annotated information about whether they were occurrences of the pronoun *ello* or substrings of antecedents or connective phrases related to these occurrences. After this, we processed the corpus to extract, for each occurrence of *ello*, information about a sequence of tokens containing the neuter pronoun, the 245 preceding tokens, and the ten following tokens. The size of the sequences extracted for each occurrence of *ello* was determined on the basis of the corpus analysis presented in Section 3.1.1 and also for the benefit of the deep learning model. The sequence length of 256 ensures that sequences usually contain the antecedent and when set this way, the data is also fully aligned in memory, making it easier and faster to read and write.

Table 1 presents a short sample of such a sequence. Note that, for brevity, no information is provided about tokens 38 837-39 024 in this sequence.

In this sample, tokens *cultivar* and *patatas* are both substrings of the ANTECEDENT of the ELLO token in position 39040. As consecutive tokens with the same class label, the antecedent of *ello* is determined to be *cultivar patatas*. Columns SEMANTIC_RELATION and TYPE_REF_EXTENT are unspecified for all tokens in the sequence, save for occurrences of the pronoun *ello*. In this example, the pronoun's antecedent is a NP[3] and the semantic relation holding between this text segment and the following discourse is ADDITIVE. The word *para* at position 39039 is the CONNECTIVE_PHRASE which, in combination with the pronoun, expresses this relation. The sequence labelling model that we present in Section 2.3.1 is finetuned using data in this format to automatically predict the values of the final three columns of Table 1 in previously unseen token sequences.

[2] `https://spacy.io/`. Last accessed 4th July 2019.

[3] In this paper, we consider gerund phrases to be noun phrases due to their distributional similarity to the latter.

Table 1 *Sample of a token sequence derived from our corpus*

Token	Position	Lemma	PoS	Semantic_ Relation	Type_Ref_ Extent	Class Label
de	38836	de	ADP	NA	NA	OTHER
...	...	...	...	...	...	...
es	39025	ser	AUX	NA	NA	OTHER
capaz	39026	capaz	ADJ	NA	NA	OTHER
de	39027	de	ADP	NA	NA	OTHER
cultivar	39028	cultivar	VERB	NA	NA	ANTECEDENT
patatas	39029	patata	NOUN	NA	NA	ANTECEDENT
,	39030	,	PUNCT	NA	NA	OTHER
pero	39031	pero	CONJ	NA	NA	OTHER
que	39032	que	SCONJ	NA	NA	OTHER
su	39033	su	DET	NA	NA	OTHER
tierra	39034	tierra	NOUN	NA	NA	OTHER
no	39035	no	ADV	NA	NA	OTHER
es	39036	ser	AUX	NA	NA	OTHER
muy	39037	muy	ADV	NA	NA	OTHER
idónea	39038	idóneo	ADJ	NA	NA	OTHER
para	39039	parir	ADP	NA	NA	CONNECTIVE_PHRASE
ello	39040	él	PRON	ADDITIVE	NP	ELLO
.	39041	.	PUNCT	NA	NA	OTHER
En	39042	En	ADP	NA	NA	OTHER
este	39043	este	DET	NA	NA	OTHER
caso	39044	casar	NOUN	NA	NA	OTHER
,	39045	,	PUNCT	NA	NA	OTHER
es	39046	ser	AUX	NA	NA	OTHER
fácil	39047	fácil	ADJ	NA	NA	OTHER
ver	39048	ver	AUX	NA	NA	OTHER
que	39049	que	SCONJ	NA	NA	OTHER
el	39050	el	DET	NA	NA	OTHER

Although *spaCy* provides information on the lemma and part of speech of each token in the sequence, this information does not contribute to the performance of our automatic sequence tagging model. When processing previously unseen sequences containing the pronoun *ello*, our tagging model (Section 2.3.1) is able to accurately predict the class labels of columns SEMANTIC_RELATION, TYPE_REF_EXTENT, and *Class Label* on the basis of the *Token* column alone.

The accuracy of a random selection method to identify the first token of the antecedent of a given occurrence of *ello* is inversely proportional to the length of the token sequence being processed, provided that the sequence actually does contain that first token. For the examples of *ello* annotated in our corpus, we tabulated the numbers of tokens occurring between the leftmost token of the antecedent and the pronoun. We found that in over 98% of cases, these leftmost tokens occurred within 100 tokens of the pronoun while in 99% of cases, they occurred within 200 tokens of the pronoun. As previously mentioned, and given the limited amount of annotated data that we had available, in an effort to exploit as much of it as possible, we conservatively set the sequence lengths in our model to 256 tokens.

From the annotated corpus described in Section 2.1, we derived 1915 sequences of the type illustrated in Table 1.

2.3 Classification and Resolution of *Ello* as a Sequence Labelling Task

The general task of pronominal anaphora resolution can be framed as a sequence labelling task in which successful methods identify a contiguous sequence of (one or more) tokens which precede the pronoun as being the antecedent of the pronoun. In the specific type of anaphora resolution that we address in this paper, the sequence labelling task is slightly more complex, with two preceding sequences needing to be identified: one comprising the antecedent of the pronoun and one occurring between the pronoun and antecedent which comprises the connective phrase expressing a semantic relation between the antecedent and the following text segment.

A number of machine learning methods have been developed to learn accurate sequence labelling models from annotated data for various tasks in NLP. These include methods based on hidden Markov models [4], maximum entropy models [21], conditional random fields [26] and recurrent neural networks integrating long short-term memory units [19].

2.3.1 Our Neural Approach to Sequence Labelling

We developed a method to automatically identify, for input sequences of tokens containing the pronoun *ello*, those tokens belonging to the antecedent of the pronoun and those tokens belonging to the connective phrase which combines with the pronoun to form a discourse marker expressing the semantic relation between the antecedent and the following text segment. Our method is also able to identify the type of semantic relation holding between the two text segments and to identify, for each example of *ello*, the type of antecedent that it takes (NP, Clause, Clause_Complex, Text_Portion).

We used BERT[4] [11], a state-of-the-art method for pre-training language representations, to build a tagging model which implements our method. BERT's model architecture is a multi-layer bidirectional Transformer encoder based on an implementation described by Vaswani et al [56]. The pretrained language model that we used provides support for 104 languages and consists of 12 layers (Transformer Blocks), 768 hidden units, 12 attention heads and has 110 million parameters.

BERT was originally designed for use in the context of transfer learning, where an initial model is derived to solve NLP tasks for which the creation of massive sets of labelled training and testing data is trivial and inexpensive. In this pretraining step, BERT was used to learn models to automatically predict missing words in cloze tests and to predict whether sentences randomly selected from the text immediately follow given test sentences. For these tasks, BERT was pretrained on over three billion words of text. From the initial parameter settings obtained through pretraining, a finetuning step can then be applied to optimise the parameter settings for new tasks. When finetuned using new hand-labelled and task-specific training data of restricted size, BERT

[4] Available at `https://github.com/google-research/bert`. Last accessed 3rd July 2019.

models have been shown to achieve great accuracy in a range of NLP tasks, including language understanding, question answering, and grounded common sense inference [11]. For our purposes, we used the corpus annotated in accordance with the scheme presented in Section 2.1.1 of this article to finetune the BERT model to identify tokens that are substrings of antecedents and connective phrases related to occurrences of the pronoun *ello*.

To predict the class labels of input tokens in our task (identifying them as substrings of ANTECEDENTs, CONNECTIVE_PHRASEs, ELLOs, or OTHER elements in input token sequences), we finetuned the BERT-Base Multilingual Cased model [11].[5] We used standard 10-fold cross validation, dividing our annotated data into 10 equal parts and successively using one part as testing data and the remaining nine as training data, to finetune the 110 million parameters of the BERT model. The settings that we used when finetuning BERT for our sequence labelling task were *max_seq_length*= 512,[6] *num_train_epochs*= 5, and *training_batch_size*= 1.[7] For our sequence labelling task, we used the set of 1915 hand-labelled token sequences described in Section 2.2 to finetune the model.

We tuned the model to simultaneously predict the values of columns SEMANTIC_RELATION, TYPE_REF_EXTENT, and *class label* in Table 1. This was achieved using a multi-task learning setup in which an additional linear, whose inputs are the shared BERT representation, generates task-specific representations to be used in the three prediction tasks.[8] After finetuning, the models were applied to the testing data to automatically predict the semantic relations, antecedent types, and class labels of tokens in the input token sequences. We refer to the implemented system as *SACRE*.[9]

One advantage of BERT models is that they include vector representations for sublexical units such as character trigrams, character bigrams, and character unigrams. These can provide a flexible treatment of out of vocabulary words. Such words can be segmented into the smallest possible number of sublexical units and representations of these can be obtained and then combined to provide an overall representation for out of vocabulary words. In the worst case, the representations of each character of an unknown word may be combined in this way to represent the word. This helps to overcome the problem observed in our use of Base$_{CRF2}$ baseline which represents word meanings through lookup of Spanish word embeddings in a large dictionary (Section 4.1).

[5] Available at `https://storage.googleapis.com/bert_models/2018_11_23/multi_cased_L-12_H-768_A-12.zip`. Last accessed 26th May 2021. Further details on the derivation of BERT's multilingual models are presented at `https://github.com/google-research/bert/blob/master/multilingual.md`. Last accessed 26th May 2021.

[6] Associating each occurrence of *ello* with a context of 512 neighbouring tokens.

[7] Tagging each sequence of 512 tokens independently of other sequences in the text.

[8] In the literature, this additional layer is usually described as being situated "on top of" the BERT layer.

[9] System to Automatically Classify and Resolve *Ello*.

3 Results

In this section, we present results of our corpus development process and results of an evaluation of our system to automatically classify occurrences of the pronoun *ello* in Spanish text and to identify the antecedents of these pronouns.

The methods described in Section 2 led to the development of two novel resources:

1. A text corpus annotated with information about 1915 occurrences of the Spanish pronoun *ello*, the antecedents of these pronouns, and the semantic relations holding between the antecedent and following text segments linked by the connective and the encapsulator;
2. Software implementing a BERT model to classify tokens in input token sequences as substrings of either:
   (a) occurrences of the pronoun *ello*,
   (b) antecedents of these pronouns,
   (c) connective phrases which, in combination with the pronoun, encode semantic relations between the antecedent and subsequent text segments, or
   (d) unrelated text.

Our results pertain to the characteristics and quality of the annotated corpus (Section 3.1) and the accuracy of our neural approach to identifying and classifying antecedents of the Spanish encapsulator *ello* and the connective phrases that combine with these pronouns to link their antecedents with subsequent text segments (Section 3.2).

3.1 The Annotated Corpus

In this section, we provide information on the characteristics of the corpus annotated in accordance with the scheme presented in Section 2.1.1. We include an analysis of inter-annotator agreement, which indicates both the reliability of the annotation and the upper limit of accuracy to be expected from automatic systems designed to replicate human performance in the automatic tagging task.

3.1.1 Corpus Analysis

We analysed the annotated corpus produced by our annotators in accordance with the annotation scheme presented in Section 2.1.1. Table 2 displays the frequency of occurrence of the annotated elements and their attributes in this corpus. According to this data, *ello* more frequently encapsulates clauses and complex clauses, and is rarely found without other linguistic elements. This illustrates *ello* as an encapsulator of complex ideas that works in combination with other linguistic particles, as stated in Section 1. CAUSAL was the

Table 2 *Frequency list of annotated elements and attributes in our corpus*

Tag	Attribute	Freq
Antecedent		1918
Connective_Phrase		1744
Ello		1916
Semantic_Relation		
	Additive	469
	Adversative	188
	Causal	1259
Type_Ref_Extent		
	Clause	1057
	Clause_Complex	715
	NP	150
	Text_Portion	2

most frequently occurring semantic relation, which serves to emphasise our observation that *ello* helps to convey complex meanings.

Table 3 displays a list of the most frequent connective phrases annotated in the corpus. The list of token sequences annotated as connective phrases has a long tail and includes 285 different phrases. Many of these are variants which differ only with respect to the punctuation marks and parentheticals tha they contain.

Table 3 *Frequency list of the ten most frequent connective phrases in our corpus (downcased)*

Connective Phrase	Freq
por	749
para	479
con	118
en lugar de	94
a pesar de	60
y	55
de	44
en vez de	33
a	29
pero	17

While the most frequently occurring connective phrases in our corpus were, at most, trigrams (Table 3), manually annotated connective phrases in our corpus were of a variety of sizes, including 6- and 7-grams such as *pero como no estamos seguros de*, *aunque no seamos más felices por*, and *a pesar de los posibles costes que*.

After the manual annotation of our corpus, cases of disagreement between annotators were resolved by an expert arbitrator. In Table 4, we provide statistics on inter-annotator agreement using the average kappa (κ) score [8] between

the four annotators and between each annotator and the final arbitrator's decision.[10]

Table 4 *Inter-annotator agreement*

Avg. between pairs of annotators	
Tag/Attribute	κ
TYPE_REF_EXTENT	0.7135
CONNECTIVE_PHRASE	0.9989
SEMANTIC_RELATION	0.6831
Avg. between each annotator and the arbitrator	
TYPE_REF_EXTENT	0.7964
CONNECTIVE_PHRASE	0.9986
SEMANTIC_RELATION	0.7964

We observed that, when considering disagreements between annotators, the four most frequent were cases in which:

- annotator 1 labelled SEMANTIC_RELATIONs as ADDITIVE when the others (including the arbitrator) labelled them as CAUSAL (295 disagreements),
- annotator 1 labelled TYPE_REF_EXTENTs as CLAUSE when the others (including the arbitrator) labelled them as CLAUSE_COMPLEX (153 disagreements),
- annotators 1, 2, and 4 labelled SEMANTIC_RELATIONs as ADDITIVE when annotator 3 and the arbitrator labelled them as CAUSAL (121 disagreements),
- annotator 2 labelled SEMANTIC_RELATIONs as CAUSAL when the others (including the arbitrator labelled them as ADDITIVE (85 disagreements).

We identify the third point of disagreement as being of most concern, as three fifths of the human participants are in disagreement with two fifths of them. However, this level of agreement is in line with other annotation exercises for NLP tasks such as anaphora resolution [30] ($0.51 \leq F_1 \leq 0.65$), including resolution of sense anaphoric pronouns [43] ($\kappa = 0.67$), discourse analysis [58] ($0.65 \leq \kappa \leq 0.85$), identification of multiword expressions (in English [44] ($\kappa = 0.79$), English and Spanish [55] ($\kappa = 0.6$ and 0.44, respectively), and Italian [54] ($\kappa = 0.65$)), and others.

The relatively high level of disagreement between annotators when marking up CAUSAL and ADDITIVE relations holding between two textual segments has a range of possible explanations. One of the most plausible is that when the two textual segments are juxtaposed (i.e. there is no explicit discourse marker or connective phrase linking the segments), the semantic relationship is implicit and governed by verbal processes. This implicitness is driven by the semantics of the main verb of the second textual segment. This verb is

[10] We used the implementation made in the *scikit-learn* machine learning library for Python to compute κ scores.

usually causal in nature (e.g. *impact*, *produce*, *generate*, *affect*, *cause*). Since there is no explicit link between the two segments connected by *ello*, it is possible that a human annotator would classify this semantic relationship of juxtaposition as ADDITIVE and not pay sufficient attention to the main causal verb of the second segment, which in fact defines the semantic relationship holding between both textual elements. Example (2) illustrates this:

(2) **Segment 1**: Los valores de las acciones ZIPER han subido progresivamente en el último trimestre. **Segment 2**: Ello **produjo** un impacto global en los mercados textiles del continente.

[**Segment 1**: The values of the ZIPER shares have progressively increased in the last quarter. **Segment 2**: This **produced** a global impact on the continent's textile markets.]

We observed that most disagreements in which CAUSAL semantic relations were identified as ADDITIVE occurred in the annotation of relations holding between juxtaposed text segments. In future annotation projects, we will highlight this phenomenon in the annotation guidelines with a view to increasing the level of agreement between annotators.

Despite the number of disagreements between annotators with regard to the annotation of SEMANTIC_RELATIONs, the figures in Table 4 indicate that inter-annotator agreement and agreement with the arbitrator range from substantial to almost perfect.[11] For this reason, we are confident that our annotation has a level of consistency and reliability sufficient to support the development of machine learning approaches for the automatic annotation of this information in Spanish texts.

3.2 Accuracy of Our Method for Automatic Classification and Resolution of *Ello*

We evaluated our sequence labelling model using standard 10-fold cross validation: dividing the data into 10 equal parts, and successively using one part as testing data, and the remaining nine as training data to finetune the BERT model, before using the finetuned models on testing data to predict the tokens' labels.

We perform 10-fold cross validation rather than splitting our annotated data into static training, development, and test portions because our dataset is relatively small and we were concerned that a single allocation of sequences to the three portions would produce a training portion that was not sufficiently representative of the entire dataset. In 10-fold cross validation, the evaluation is performed over ten sequence labelling models which, altogether, are derived from the entire dataset.

We present the results of two sets of evaluation metrics:

[11] According to the scale proposed by Viera and Garrett [57].

1. Precision (P), recall (R), and F_1-score: In our evaluation, for each distinct class label, true positives (TP) are tokens with that class label in the gold standard which our method correctly predicts as having that class label, false positives (FP) are tokens that do not have that class label in the gold standard which our method mistakenly predicts as having that class label, and false negatives (FP) are tokens with that class label in the gold standard which our method mistakenly predicts as having a different class label. Then
   $P = \frac{TP}{(TP+FP)}$, $R = \frac{TP}{(TP+FN)}$, and $F_1 = \frac{2 \times P \times R}{P+R}$.
2. First token accuracy for antecedents. In this context, we use two metrics:

   – *Correct token* is the proportion of tokens predicted by our method as having the class label ANTECEDENT and being the first in a single- or multi-token antecedent identified by our method that is also the first token in a single- or multi-token antecedent in the gold standard. The class labels predicted by Method 1 in Table 5 make token T4 (in column *Pred. class label (Method 1)*) an example of such a token.
   – *Within 1 token* is the proportion of tokens for which these conditions hold but also awarding cases where the token predicted by our model occurs within one token of the first token in a single- or multi-token antecedent in the gold standard. The class labels predicted by Method 2 in Table 5 make token T3 (in column *Pred. class label (Method 2)*) an example of such a token as it occurs within one token of the true first token of the antecedent (token T4).[12]

Table 5 True class labels and class labels predicted by three methods for tokens in a hypothetical sequence

Token	True class label	Pred. class label (Method 1)	Pred. class label (Method 2)	Pred. class label (Method 3)
T1	NA	NA	NA	NA
T2	NA	NA	NA	ANTECEDENT
T3	NA	NA	ANTECEDENT	ANTECEDENT
T4	ANTECEDENT	ANTECEDENT	ANTECEDENT	ANTECEDENT
T5	ANTECEDENT	ANTECEDENT	ANTECEDENT	NA
T6	ANTECEDENT	ANTECEDENT	ANTECEDENT	NA
T7	NA	NA	NA	NA
T8	NA	NA	NA	NA

These metrics indicate the accuracy of our method in locating the start point of the token sequences which contain ANTECEDENT, CONNECTIVE_PHRASE, and ELLO elements.

Table 6 presents mean scores and standard deviations for precision (P), recall (R), and F_1-scores for ANTECEDENT, CONNECTIVE_PHRASE, ELLO, and

[12] By contrast, token T2 in column *Pred. class label (Method 3)* is not of this type because it is two tokens away from the true start of the ANTECEDENT.

OTHER tokens. It also displays mean scores and standard deviations for identification of the correct first tokens in ANTECEDENTs, and within 1 token accuracy for identification of these tokens over the ten folds of evaluation. This latter metric shows that for almost half of the occurrences of *ello*, SACRE is able to identify the first word of its antecedent within one token.

Table 6 *Results for automatic labelling of sequences consisting of antecedents, connective phrases, and occurrences of ello*

	ANTECEDENT		CONNECTIVE_PHRASE		ELLO		OTHER	
	Mean	σ	Mean	σ	Mean	σ	Mean	σ
Precision	0.77	0.06	0.77	0.19	0.997	0.003	0.97	0.01
Recall	0.76	0.06	0.72	0.14	0.994	0.011	0.97	0.006
F_1-score	0.76	0.02	0.73	0.14	0.99	0.006	0.97	0.005
First token accuracy for antecedents								
Correct token	0.43	0.04						
Within 1 token	0.48	0.05						

Tables 7 and 8, respectively, present figures for the accuracy with which our model (Section 2.3.1) identifies the semantic relations expressed by the complex discourse markers linking antecedent and subsequent text segments in our corpus and the accuracy with which our model identifies, for each occurrence of *ello*, the type of antecedent that it encapsulates.

Table 7 *Results for classification of the semantic relations expressed by complex discourse markers (connective phrases combined with occurrences of ello)*

Semantic Relation		Recall	Precision	F_1-score
NA	Mean	1.00	1.00	1.00
	σ	0.00	0.00	0.00
ADDITIVE	Mean	0.11	0.38	0.14
	σ	0.09	0.23	0.07
ADVERSATIVE	Mean	0.70	0.82	0.74
	σ	0.22	0.12	0.17
CAUSAL	Mean	0.90	0.70	0.78
	σ	0.11	0.10	0.07

4 Discussion

4.1 Significance of Our Results

The development of a new annotated corpus of Spanish texts, annotated with information about extant examples of the pronoun *ello*, the antecedents that

Table 8 *Results for classification of antecedent types*

Type		Recall	Precision	F_1-score
NA	Mean	1.00	1.00	1.00
	σ	0.00	0.00	0.00
CLAUSE	Mean	0.63	0.66	0.63
	σ	0.05	0.15	0.09
CLAUSE_COMPLEX	Mean	0.66	0.54	0.57
	σ	0.10	0.14	0.07
NP	Mean	0.17	-	-
	σ	0.17	-	-
TEXT_PORTION	Mean	-	-	-
	σ	-	-	-

they encapsulate, and the semantic relations between the antecedent and subsequent text segments is significant. To date, there has been little work in NLP concerned with the development of methods to resolve this type of anaphora. Part of the reason for this is the current lack of annotated resources to facilitate evaluation of these methods.

In developing this annotated resource together with a practical system to resolve this type of anaphora and to detect the semantic relations between linked text segments, the research described in this article makes a significant advance on the state of the art of NLP for Spanish. In Section 4.1.1, we also show that our SACRE system is significantly more accurate than three baseline systems and sets a strong benchmark for resolution of the pronoun *ello* and identification of the connective phrases linking antecedent and subsequent text segments. In Section 4.1.2, we show that our method outperforms a majority class baseline in the task of classifying occurrences of the pronoun *ello* with respect to the syntactic type of the antecedent that it encapsulates and the semantic relations between the two text segments.

4.1.1 Identifying Antecedent *and* Connective_Phrase *Token Sequences*

We implemented three baseline methods to identify tokens that are substrings of the antecedents and connective phrases related to occurrences of the pronoun *ello*. The first, Base$_{PS}$, identifies the sentence preceding the sentence which contains the pronoun as the antecedent of that pronoun. That is, every token in the preceding sentence is tagged ANTECEDENT. Every token between the rightmost token tagged as ANTECEDENT and the token tagged ELLO is tagged as CONNECTIVE_PHRASE.

In previous work, CRF classifiers have been shown to be effective in a variety of linguistic sequence labelling tasks including named entity recognition [29], shallow parsing [48], clause boundary identification [27], the automatic identification of compound noun phrases and their conjoins [49], and syntactic

constituent analysis [16]. Our second baseline, Base$_{CRF1}$ uses a sequence based tagging model exploiting conditional random fields (CRF) [26,52] to tag tokens that are substrings of ANTECEDENT, CONNECTIVE_PHRASE, ELLO, and OTHER elements. We used the CRF++ package [25] to derive this sequence tagging model.

The Base$_{CRF1}$ model exploits information about the lemmas and parts of speech of tokens and input token sequences and uses feature templates providing information about neighbouring tokens to condition the probability of each token being of a particular class. For this baseline, we specified feature templates to provide information about the words, lemmas, and parts of speech of the two tokens preceding the token to be classified, the token itself, and the two tokens following this token. We also included feature templates to exploit information about word, lemma, and part of speech bigrams involving the token to be classified. This included bigrams immediately preceding the token, bigrams of which the token is the second element, bigrams of which the token is the first element, and bigrams immediately following the token. The final set of feature templates encoded information about part of speech trigrams in which the token to be classified occurs.

The Base$_{CRF2}$ model is an expansion of the Base$_{CRF1}$ model which includes Spanish word embeddings (300 features) in the representations of tokens being classified. Our motivation for using Spanish word embeddings was an intuition that identification of the antecedent of the pronoun *ello* would be facilitated by access to semantic information about tokens and token sequences. We used the publicly available Spanish word embeddings in the SBW dataset.[13] The CRF tagger exploited this information via a set of 300 unigram feature templates. With 32GB RAM, we lacked the computational resources needed to derive tagging models using more sophisticated feature templates based on the word embeddings of token bigrams and trigrams.

Table 9 presents the accuracy scores achieved by these baseline methods (Columns $Base_{PS}$, $Base_{CRF1}$, and $Base_{CRF2}$) and by our new model (*SACRE*). We observe that inclusion of word embeddings in the token representation used by $Base_{CRF2}$ did not bring about improvements in F_1-score in our sequence labelling task. Inspection of our data revealed that 15.98% of token lemmas (23.34% of word types) occurring in these sequences were out of vocabulary due to typographical irregularities in the input texts. $Base_{CRF2}$ relies on exact matching of word lemmas and cannot fall back on sublexical matching, as the SACRE system can. For this reason, the $Base_{CRF2}$ method is unable to associate every token in the input test sequences with semantic information from Spanish word embeddings. This has an adverse effect on the accuracy of $Base_{CRF2}$.

In the task of identifying ANTECEDENT and CONNECTIVE_PHRASE token sequences, SACRE is more accurate than any of the three baselines. It is also

[13] Available from `http://cs.famaf.unc.edu.ar/~ccardellino/SBWCE/SBW-vectors-300-min5.bin.gz`. Last accessed 22nd August 2019. These word embeddings were derived from the Spanish Billion Word corpus, available from `http://crscardellino.github.io/SBWCE/`. Last accessed 22nd August 2019.

Table 9 *Baseline results for automatic labelling of sequences consisting of antecedents, connective phrases, and occurrences of ello*

Class label	F_1 (Avg)			
	Base_{PS}	Base_{CRF1}	Base_{CRF2}	SACRE
ANTECEDENT	0.44	0.20	0.19	0.76
CONNECTIVE_PHRASE	0.08	0.68	0.66	0.73
ELLO	0.97	0.97	0.97	0.99
OTHER	0.93	0.95	0.94	0.97

more accurate in its identification of token sequences identified by our annotators as occurrences of *ello* and tokens classed as OTHER. As noted earlier, typographical issues and the accidental inclusion of adjacent punctuation marks in ELLO elements by our annotators means that the automatic identification of these elements is not 100% accurate.

When discussing statistically significant differences in the accuracy of two systems, we will use the terms *significant* and *insignificant* as shorthand for statistically significant/insignificant, respectively. Except where noted, the Bonferroni corrected significance level $\alpha = 0.016$ for pairwise comparisons among four systems.[14] We applied independent t-tests over all folds of the ten-fold cross-validation to test for significant differences between evaluation results obtained by each system.

We compared precision and recall of the four systems when identifying tokens in each of four classes (ANTECEDENT, CONNECTIVE_PHRASE, ELLO, and OTHER), for a total of 24 comparisons. In total, SACRE was the significantly superior system in twelve of these comparisons, while $Base_{PS}$, $Base_{CRF1}$, and $Base_{CRF2}$ were each superior in two. In the comparisons, we found:

- The SACRE system obtains significantly greater F_1-scores than any of the other baseline methods when classifying tokens, regardless of their class label. The differences in F_1-score obtained by SACRE and the other baselines when identifying tokens of the ELLO class are statistically significant ($p < 0.002$). This observation also holds for comparisons involving the identification of tokens of the OTHER class ($p < 0.001$ in all cases).
- When identifying ANTECEDENT tokens, Base_{PS} is significantly more accurate than either of the two baselines exploiting CRF models ($p \ll 0.016$ in both cases).
- When identifying CONNECTIVE_PHRASE tokens, the two baselines using CRF models (Base_{CRF1} and Base_{CRF2}) obtained significantly greater F_1-scores than Base_{PS} ($p \ll 0.016$). This observation also holds when identifying OTHER tokens, which comprise the overwhelming majority in the dataset ($p < 0.0009$, in both cases).

[14] Adjusted from $\alpha = 0.05$ for comparisons between two systems.

4.1.2 Classifying Ello with Regard to Type_Ref_Extent *and* Semantic_Relation

We developed a majority class baseline to classify the antecedent type of each pronoun *ello* and the semantic relations between the antecedent and subsequent text segments. The majority class for antecedent type was CLAUSE while the majority class for semantic relation was CAUSAL. As a result, the majority class baseline achieves F_1-scores of 0.55 and 0.66 for identification of antecedent type and semantic relation, respectively.

Our classification methods, based on the model presented in Section 2.3.1, compare favourably with the majority class baseline (Tables 7 and 8).

5 Conclusions

In this paper, we presented the development of a new annotated corpus of Spanish text which encodes information about occurrences of the encapsulating pronoun *ello*, antecedents of these encapsulators, and the semantic relations expressed by complex discourse markers containing the pronoun which hold between antecedent and subsequent text segments.

Our analysis of the corpus revealed that in the majority of cases, the antecedent of *ello* is a single clause, though the next most frequent type of antecedent is that of clause complex, consisting of multiple clauses in contiguous text. The semantic relations between the antecedent and subsequent text segments are usually causal or additive and expressed by connective phrases consisting of the prepositions *por* or *para*. Our corpus was annotated by 4 annotators and assessment of inter-annotator agreement revealed substantial to perfect agreement, indicating that the annotation is consistent and reliable.

We used the annotated corpus to train an automatic sequence tagging approach to perform a novel and challenging anaphora resolution task: the automatic identification of antecedents of occurrences of *ello*. Our neural approach exploits part of speech tagging and finetuned BERT embeddings to predict, for each token in the 245 tokens preceding an occurrence of *ello*, those which are part of the antecedent and those which are part of the connective phrase which, in combination with the pronoun, link the antecedent and subsequent text segments. In addition to identifying these tokens, our neural sequence tagging method also classifies the tokens to identify the semantic relations between the two text segments and the grammatical type of antecedent that the pronoun has.

Evaluation of our automatic method revealed that, with accuracy scores exceeding 70% for every aspect of the task, it is quite reliable. The accuracy obtained by our method compares favorably with that of other automatic methods for the challenging NLP task of anaphora resolution.

In our experiments, we annotated a corpus and derived sequence tagging models from textbooks about Economics written in Spanish. In future work, it will be interesting to expand our annotated corpus to texts of other domains

and registers and to investigate the applicability of models derived from one domain to input text from another domain. The availability of additional data of this type would also provide us with the opportunity to examine potential benefits brought by training our models on larger training sets consisting of texts from multiple domains.

References

1. Ariel, M.: Referring and accessibility. Journal of Linguistics **64**, 65–87 (1988)
2. Ariel, M.: The function of accessibility in a theory of grammar. Journal of Pragmatics **16**, 443–463 (1991)
3. Ariel, M.: Cognitive universals and linguistic conventions. the case of resumptive pronouns. Studies in Language **23**, 217–269 (1999)
4. Baum, L.E., Petrie, T.: Statistical inference for probabilistic functions of finite state markov chains. The Annals of Mathematical Statistics **37**(6), 1554–1563 (1966). DOI 10.1214/aoms/1177699147. URL `https://doi.org/10.1214/aoms/1177699147`
5. Bello, A.: Gramática de la Lengua Castellana. Roger & Chernovitz Editores, Paris (1911)
6. Benveniste, E.: Problemas de Lingüística General. Tono I. Siglo XXI Editores, México (1980)
7. Borreguero, M.: Naturaleza y función de los encapsuladores en los textos informativamente densos (la noticia periodística). Cuadernos de Filología Italiana **13**, 73–95 (2006)
8. Cohen, J.: A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement **20**(1), 37–46 (1960)
9. Cornish, F.: Anaphora, Discourse, and Understanding. Oxford University Press, Oxford (1999)
10. Cornish, F.: How indexicals function in texts: Discourse, text, and one neo-Gricean account of indexical reference. Journal of Pragmatics **40**(6) (2008)
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). URL `https://www.aclweb.org/anthology/N19-1423`
12. van Dijk, T.A., Kintsch, W.: Strategies of Discourse Comprehension. Academic Press, New York (1983)
13. Fernández, O.: El pronombre personal. Formas y distribuciones. Pronombre átonos y tónicos. In: I. Bosque, V.C. Demonte (eds.) Gramática Descriptiva de la Lengua Española, pp. 1209–1273. Espasa Calpe, Madrid (1999)
14. Figueras, C.: La jerarquía de accesibilidad de las expresiones referenciales en español. Revista Española de Lingüística **32**, 53–96 (2002)
15. Francis, N.: Anaphoric Nouns. University of Birmingham, Birmingham (1986)
16. Gómez-Rodríguez, C., Vilares, D.: Constituent parsing as sequence labeling. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 1314–1324. Association for Computational Linguistics (2018). URL `http://aclweb.org/anthology/D18-1162`
17. González-Ruiz, R.: Algunas notas en torno a un mecanismo de cohesión textual: La anáfora conceptual. In: Estudios sobre el Texto. Nuevos Enfoques y Propuestas, pp. 247–278. Peter Lang, Frankfurt (2009)
18. Halliday, M., Hasan, R.: Cohesion in English. Longman, London (1976)
19. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computing **9**(8), 1735–1780 (1997). DOI 10.1162/neco.1997.9.8.1735. URL `http://dx.doi.org/10.1162/neco.1997.9.8.1735`

20. Honnibal, M., Johnson, M.: An improved non-monotonic transition system for dependency parsing. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1373–1378. Association for Computational Linguistics, Lisbon, Portugal (2015). URL https://aclweb.org/anthology/D/D15/D15-1162
21. Jaynes, E.T.: Information theory and statistical mechanics. Physical Review **106**(4), 620–630 (1957). DOI 10.1103/PhysRev.106.620. URL http://prola.aps.org/abstract/PR/v106/i4/p620_1
22. Kennison, S.: Comprehending the pronouns her, him, and his: Implications for theories of referential processing. Journal of Memory and Language **49**(3), 335–352 (2003)
23. Kennison, S., Trofe, J.: Comprehending pronouns: A role for word-specific gender stereotype information. Journal of Psycholinguistic Research **32**(3), 355–378 (2003)
24. Kintsch, W.: Comprehension a Paradigm for Cognition. Academic Press, New York (1998)
25. Kudo, T.: CRF++: Yet another CRF toolkit. Software available at http://crfpp. sourceforge. net (2005)
26. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning, pp. 282–289. Morgan Kaufmann (2001)
27. Lakshmi, S., Ram, R.V.S., Sobha, L.D.: Clause Boundary Identification for Malayalam Using CRF. In: Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages (MTPIL-2012), pp. 83–92. Association for Computational Linguistics, Mumbai, India (2012)
28. López Samaniego, A.: La categorización de entidades del discurso en la escritura profesional. Phd thesis, Universitat de Barcelona, Barcelona, España (2011)
29. McCallum, A., Li, W.: Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4, pp. 188–191. Association for Computational Linguistics (2003)
30. Mitkov, R., Evans, R., Orasan, C., Barbu, C., Jones, L., Sotirova, V.: Coreference and anaphora: Developing annotating tools annotated resources and annotation strategies. In: Proceedings of DAARC-2000, UK, pp. 49–58 (2000)
31. Montolío, E.: Construcciones conectivas que encapsulan. [A pesar de + SN] y la escritura experta. Cuadernos AISPI **2**, 115–132 (2013)
32. Montolío, E.: Mecanismos de cohesión (II). Los conectores. In: E.M. (Dir) (ed.) Manual de Escritura Académica y Profesional, pp. 9–92. Ariel, Barcelona (2014)
33. Orăsan, C.: PALinkA: a highly customizable tool for discourse annotation. In: Proceedings of the 4th SIGdial Workshop on Discourse and Dialog, pp. 39 – 43. Sapporo, Japan (2003). URL http://clg.wlv.ac.uk/papers/palinka-final.pdf
34. Parodi, G.: Comprensión de Textos Escritos. Teoría de la Comunicabilidad. Eudeba, Buenos Aires (2014)
35. Parodi, G., Burdiles, G.: Encapsulación y tipos de coherencia referencial y relacional: el pronombre "ello" como mecanismo encapsulador en el discurso escrito de la economía. Onomázein **33**(1), 107–129 (2016)
36. Parodi, G., Burdiles, G.: Los pronombres neutros 'esto', 'eso' y 'aquello' como mecanismos encapsuladores: coherencia referencial y relacional. Spanish in Context **16**(1), 104–127 (2019)
37. Parodi, G., Julio, C., Nadal, L., Burdiles, G., Cruz, A.: Always look back: Eye movements as a reflection of anaphoric encapsulation in Spanish while reading the neuter pronoun ello. Journal of Pragmatics (132), 47–58 (2018)
38. Parodi, G., Julio, C., Nadal, L., Cruz, A., Burdiles, G.: Stepping back to look ahead: Neuter encapsulation and referent extension in counter-argumentative and causal semantic relations in spanish. Language & Cognition **11**(3), 431–454 (2019)
39. Portolés, J.: Pragmática para Hispanistas. Longman (2004)
40. Prandi, M.: The Building Blocks of Meaning. Benjamins, Amsterdam/Philadelphia (2004)
41. RAE: Diccionario Panhispánico de Dudas. Santillana, Bogotá (2005)
42. RAE & ASALE: Nueva Gramática de la Lengua Española. Manual [New grammar of Spanish language. Handbook]. Espasa, Buenos Aires (2010)

43. Recasens, M., Hu, Z., Rhinehart, O.: Sense Anaphoric Pronouns: Am I One? In: Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016), pp. 1–6. Association for Computational Linguistics, San Diego, California (2016). DOI 10.18653/v1/W16-0701. URL https://www.aclweb.org/anthology/W16-0701
44. Rohanian, O., Taslimipoor, S., Yaneva, V., Ha, L.A.: Using gaze data to predict multiword expressions. In: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, pp. 601–609. INCOMA Ltd., Varna, Bulgaria (2017). DOI 10.26615/978-954-452-049-6_078. URL https://doi.org/10.26615/978-954-452-049-6_078
45. Sanders, T., Spooren, W., Noordman, L.: Toward a taxonomy of coherence relations. Discourse Processes pp. 1–35 (1992). DOI 10.1080/01638539209544800
46. Sanders, T., Spooren, W., Noordman, L.: Coherence relations in a cognitive theory of discourse representation. Cognitive Linguistics pp. 93–134 (1993). DOI 10.1515/cogl.1993.4.2.93
47. Schmid, H.: English Abstract Nouns as Conceptual Shells: From Corpus to Cognition. de Gruyter, Berlin-New York (2000)
48. Sha, F., Pereira, F.: Shallow parsing with conditional random fields. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, pp. 134–141. Association for Computational Linguistics (2003)
49. Shimbo, M., Hara, K.: A discriminative learning model for coordinate conjunctions. In: Proceedings of the 2007 Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 610–619. Prague (2007)
50. Sinclair, J.: Written discourse structure. In: J. Sinclair, M. Hoey, G. Fox (eds.) Techniques of Description. Spoken and Written Discourse, pp. 6–31. Routledge, London (1993)
51. Sinclair, J.: Trust the text. In: M. Coulthard (ed.) Advances in Written Text Analysis, pp. 6–31. Routledge, London (1994)
52. Sutton, C., McCallum, A.: An introduction to conditional random fields. Foundations and Trends in Machine Learning **4:4**, 268–373 (2011)
53. Tadros, A.: Predictive categories in expository texts. In: M. Coulthard (ed.) Advances in Written Text Analysis, pp. 69–82. Routledge, London (1994)
54. Taslimipoor, S., Desantis, A., Cherchi, M., Mitkov, R., Monti, J.: Language resources for italian: towards the development of a corpus of annotated italian multiword expressions. In: P. Basile, A. Corazza, F. Cutugno, S. Montemagni, M. Nissim, V. Patti, G. Semeraro, R. Sprugnoli (eds.) Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), CEUR Workshop Proceedings, vol. 1749. Napoli, Italy (2016)
55. Taslimipoor, S., Mitkov, R.: Computational phraseology light: automatic translation of multiword expressions without translation resources. Yearbook of Phraseology **7**(1), 149–166 (2016)
56. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., ukasz Kaiser, L., Polosukhin, I.: Attention is all you need. In: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (eds.) Advances in Neural Information Processing Systems 30, pp. 5998–6008. Curran Associates, Inc. (2017). URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf
57. Viera, A.J., Garrett, J.M.: Understanding interobserver agreement: The kappa statistic. Family Medicine **37**(5), 360–363 (2005)
58. Wang, X., Bruno, J., Molloy, H., Evanini, K., Zechner, K.: Discourse Annotation of Non-native Spontaneous Spoken Responses Using the Rhetorical Structure Theory Framework. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 263–268. Association for Computational Linguistics, Vancouver, Canada (2017)
59. Zulaica, I., Gutiérrez, J.: Hacia una semántica computacional de las anáforas demostrativas. Linguamática **1**(2), 81–90 (2009)