Grammatical Modeling of a Nominal Ellipsis Grammar for Spanish

Hazel Barahona(✉) and Walter Koza(✉)

Pontificia Universidad Católica de Valparaíso-Project FONDECyT, 1171033 Valparaíso, Chile
hazel.barahona.g@mail.pucv.cl, walter.koza@pucv.cl

Abstract. The objective of this work is to formalize nominal ellipsis in Spanish – a grammatical mechanism in which one element is silenced with its syntactic structure, under certain syntactic restrictions – through the creation of an algorithm that automatically recognizes and replaces elided elements in natural language texts. Based on the proposal by Saab [1, 2], this paper proposes a series of formalisms represented in NooJ [3, 4] to grammatically model this phenomenon. The methodology was tested on journalistic texts extracted from the Internet. The results obtained (100% precision; 82% coverage; 90.10% F-measure) show that the developed algorithm is useful for the recognition of ellipses.

Keywords: Nominal ellipsis · Generative grammar · Automatic identification · NooJ · Spanish Language

1 Introduction

This work describes a computational modeling of nominal ellipsis in Spanish. We begin with the theoretical assumptions of generative grammar following studies by Merchant [5]; Sag and Hankamer [6]; Culicover and Jackendoff [7]; and particularly, Saab [1, 2]. Specifically, the paper describes an algorithm developed for the automatic recognition of ellipses and the replacement of the elided element in natural language texts. Ellipsis consists of a grammatical mechanism avoidant of lexical redundancy [8], in which an element of the clause is not pronounced – i.e., it is elided – under certain syntactic restrictions. In this regard, Saab [1] and Merchant [9] point out that, in this phenomenon, the silenced element is recognized through a structural identity relationship with a preceding element that allows for the identification of said element and its position inside the clause.

This phenomenon has been addressed from different perspectives, including syntactic [5, 10–14] and semantic [6, 9] approaches, as well as within the frameworks of Head-Driven Phrase Structure Grammar [15] and Simpler Syntax [16]. Regarding research linked to computational linguistics, the works of Hardt [17]; Mitkov [18]; Nielsen [19]; Rello, Baeza-Yates, and Mitkov [20]; and McShane and Babkin [21] have undertaken analyses of ellipsis via corpus-labeling methodologies that identify and retrieve the elided element. However, those proposals tend not to include implicit syntactic mechanisms. To remedy this, this work seeks to formalize nominal ellipsis based on the processing

B. Bekavac et al. (Eds.): NooJ 2020, CCIS 1389, pp. 85–95, 2021.
https://doi.org/10.1007/978-3-030-70629-6_8

of morphosyntactic information. The contribution to the literature is double: on the one hand, we provide an efficient algorithm for applied computational linguistics tasks; and on the other, we develop a tool to evaluate the reach of theoretical generative studies as applied to nominal ellipsis.

For such purposes, we propose a formal description of ellipsis, modeled using an electronic dictionary under NooJ syntactic grammars, and tested on journalistic texts extracted from the Internet. The results obtained (100% precision; 82% coverage; 90.10% F-measure) show that the developed algorithm is useful for the recognition of ellipsis in natural language texts, while also showing that the generative theoretical proposals are adequate for the analysis of this phenomenon.

The article is organized as follows: first, we present the general context of ellipsis; second, we discuss the concepts relevant to the theory of Identity [1, 2], with special focus on aspects relevant to ellipsis; third, we describe the data processing tasks that we performed, considering the syntactic structure of determiner phrases (DP) in Spanish; and fourth, we present our results and conclusions.

2 Ellipsis: Definition and Types

Ellipsis is a syntactic mechanism through which, under certain structural particularities, an element is not pronounced (that is, it is elided) [1, 2, 5, 22]. In this regard, two types can be recognized: nominal ellipsis, in which a noun is elided (1a); and verbal ellipsis, in which a verb is elided (1b):

(1) a. *El perro de mi madre y el ~~perro~~ del vecino mordieron al periodista*
[my mother's dog and the neighbor's ~~dog~~ bit the journalist.]

b. *Juan lee un libro y María ~~lee~~ una revista*
[Juan is reading a book and María ~~is reading~~ a magazine.]

This mechanism acts at different syntactic levels. Thus, Spanish nominal ellipsis is a nuclear ellipsis of a DP; and, in Spanish verbal ellipsis, mechanisms of sentence ellipsis are involved. It is important to clarify that nominal ellipses do not only occur between coordinating DPs (2a), but also as part of a sentence ellipsis (2b-c). Here, the verbal assembling domains expand the locality between the preceding element (in bold) and its place in the phrase in which it is elided (strikethrough). In (3), the types of verbal ellipsis are shown. Since the objective of this work is not to address verbal ellipsis, only the verbal types that appear in the different studies are mentioned.

(2) a. *La guitarra acústica y la ~~guitarra~~ eléctrica*
[the acoustic guitar and the electric ~~guitar~~.]

b. *Los **niveles** de colesterol disminuyen y los ~~niveles~~ de azúcar aumentan*
[cholesterol **levels** decrease and sugar ~~levels~~ increase.]

c. *El **presidente** de Rusia homenajeó al ~~presidente~~ de China*
[Russia's **president** paid homage to China's ~~president~~.]

d. *Las **embarazadas** que asisten a terapia y las ~~embarazadas~~ que no asisten fueron internadas en el centro médico*
[**pregnant women** who go to therapy and ~~pregnant women~~ who do not were hospitalized in the medical center.]

(3) a. *Antonio **toca** el violín y Alex ~~toca~~ el chelo*
[Antonio **plays** the violin and Alex ~~plays~~ the cello.]

b. ***Antonio viajó** a Chile en enero y ~~Antonio viajó~~ a Brasil en agosto*
[**Antonio travelled** to Chile in January and ~~Antonio travelled~~ to Brasil in August.]

c. *Antonio **gana mucho dinero** y María también ~~gana mucho dinero~~*
[Antonio **makes a lot of money** and Maria also ~~makes a lot of Money~~.]

d. *Antonio intentó entrar **al edificio**, pero no pudo ~~entrar al edificio~~*
[Antonio tried to enter **the building**, but he could not ~~enter the building~~.]

e. ***Antonio golpeó** a alguien, pero no sé a quién ~~golpeó Antonio~~*
[**Antonio punched** someone, but I do not know whom ~~Antonio punched~~.]

f. *– ¿Quién llamó?* [who called?]
–Antonio.

One of the theories that best explains nominal (2) and verbal (3) ellipses in Spanish is that of Saab [1, 2], which is the basis for Distributed Morphology [DM, 23]. According to this theory, ellipsis is defined as a non-insertion mechanism of lexical features in which, during derivation, a particular feature, denominated as [I], blocks the insertion of phonological features. The allocation of this feature [I] is the product of a transformational operation called *Identity*, whose result is the non-pronunciation – i.e., silencing – of the elements that intervene in the syntactic operation. Saab [1] formally defines the ellipsis domain as follows:

(4) a. An abstract morpheme α is identical to the abstract morpheme β if and only if α and β match all their morphosyntactic and semantic features.
b. An A root is identical to a B root if and only if A and B share the same index.

This definition (4) establishes three contexts in which ellipsis intervenes: (i) cases of partial identity with grammatical results; (ii) cases of partial identity with ungrammatical results; and (iii) cases in which the identity is total, but the result is ungrammatical. This

work considered only the first two, which correspond to the features of number (5) and gender (6) in Spanish, respectively.

(5) a. Antonio quiere más a sus gatos que al gato de Pedro
[Antonio loves his own cats more than Pedro's cat.]

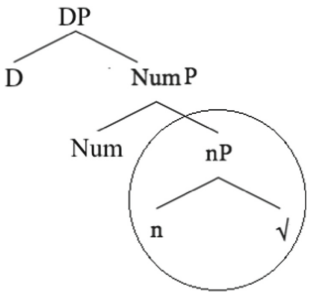
b. *Antonio quiere más a su gato que a los gatos de Pedro*
[Antonio loves his own cat more than Pedro's cats.]

(6) a. **Antonio quiere más a su gato que a la gata de Pedro*
[Antonio loves his own cat more than Pedro's [female] cat.]

b. *Antonio quiere más a sus gatas que al gato de Pedro*
[Antonio loves his [female] cats more than Pedro's cat.]

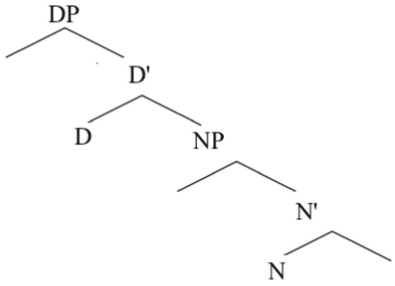
For Saab [1, 2], the ungrammaticalness of the examples in (6) is caused by the fact that, in Spanish, gender features are a property of the morphological root, in which the root is attached to the gender feature first, and then to the number feature. Therefore, for that author, number is a functional category that does not intervene in the domain of nominal ellipsis in Spanish, such as in (7).

(7)



3 The Formalization of Syntactic Information

This work adopts Abney's [24] *DP hypothesis*, in which the determiner is a functional category that can take a noun phrase as a complement; namely, it is possible to represent a structure of constituents with a particular internal structure such as the one in (8).



In Spanish, the determiner occupies the highest position, which allows the ellipsis of N in the lowest position. Furthermore, structurally speaking, the determiner is an important feature as the nucleus of the phrase, which is why it may not be an optional element (9a and 10b); and why it is also a mechanism for the recognition of the elided element, since it must agree with the elided N. In (10), the elided element is indicated with strikethrough text.

(8) a. **guitarra roja*
[red guitar]
b. *La guitarra roja*
[the red guitar]

(9) a. *La guitarra roja y la ~~guitarra~~ eléctrica*
[the red guitar and the electric ~~guitar~~.]

Spanish data was analyzed under a constituent organization to outline the determiner phrase (8), resulting in three types of syntactic structures with nuclear ellipses (N). The first is that of *coordination between determiner phrases* (11a); the second, *coordination of clauses* (12a); and the third, *predicative argument* (13a). These three syntactic structures were formalized as shown respectively in pairs (11b), (12b), and (13c). The elided element is indicated with strikethrough text.

(10) a. *El hijo de Juan y el ~~hijo~~ de Pedro*
[Juan's son and Pedro's ~~son~~].

b. [[PRENOM]$_1$ **N** [POSTNOM]$_1$]**SDI Coord** [[PRENOM]$_2$ **N** [POSTNOM]$_2$]SDII

(11) a. *El hijo de Juan fue al cine y el ~~hijo~~ de Pedro fue al teatro*
[Juan's son went to the cinema and Pedro's ~~son~~ went to the theater.]

b. [[PRENOM]$_1$ **N** [POSTNOM]$_1$]SDI **[V + ARG (ADJ)]$_1$ Coord** [[PRENOM]$_2$ **N [POSTNOM]$_2$]**SDII [V + ARG (ADJ)]$_2$

(12) a. *El presidente de Francia homenajeó al presidente de Perú*
[the president of France paid homage to the president of Peru.]

b. [[[PRENOM]$_1$ **N** [POSTNOM]$_1$]SDI]_**arg0 PRED** [[[PRENOM]$_1$ **N** [POSTNOM]$_1$]SDI]_arg1

These structures explain the phenomenon of ellipsis as a mechanism presenting under two procedures: elision (in the case of production); and identification (in the case of comprehension).

3.1 Rules for Ellipsis: Elision and Identification

The process of *elision* defines the production of a nominal ellipsis, given its articulation of the syntactic and surface structures of Spanish (14).

(13) a. Given a coordination AˆB, where A and B are DPs and B has a structure equal to A with a root object as an NP nucleus equal to that of A, elide the NP nucleus object of B.
b. Given a Predicate-Argument Structure (PAS) A, with arguments α, β, γ…, where α has a structure equal to that of the argument to its right and which has an NP whose root nucleus is equal to the NP in α, elide the nucleus of such NP.

On the other hand, *identification* describes the comprehension of the elliptical phenomenon, specifically, how the syntactic elements missing in the grammatical structure are computationally recognized (15).

(14) a. Given a coordination AˆB, where B is a DP equal to A, and the NP nucleus is elided:

- Copy the nucleus root and the NP gender object of A
- Copy the Det number of B

b. Given a PAS A: P (α, β y γ), with an object to the right of α that represents the elision of its NP nucleus:

- Copy the root of the NP object of α
- Copy the determiner number of β or γ (according to the elided object)

This work computationally modeled (14) and (15). The creation of resources in NooJ is described below.

3.2 Computational Modeling in NooJ

The computational implementation used a general Spanish language dictionary [25] containing 72,593 entries. Lemmas corresponding to nouns, adjectives, and verbs were associated with inflected model grammars.

For the identification and replacement of the elided element in the nominal ellipsis, the following syntactic grammar was created.

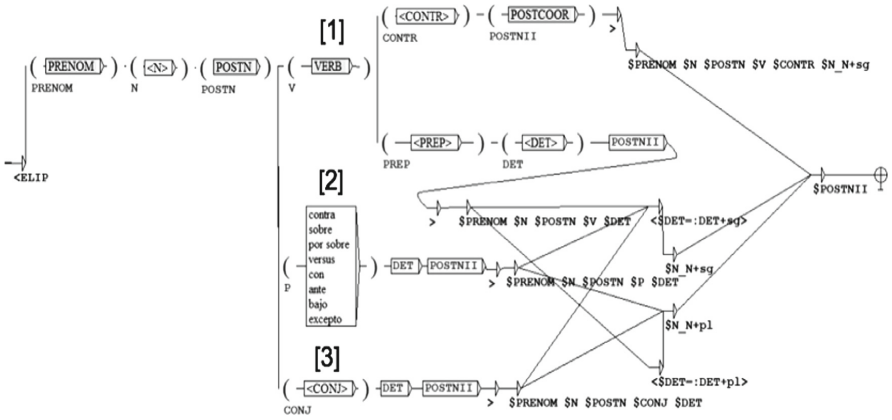


Fig. 1. Grammar for the identification and replacement of nominal ellipsis in Spanish.

The numbers in Fig. 1 indicate the types of ellipsis found in the corpus: with [1], cases of predication; with [2] and [3], coordination between determiner phrases.

The variables represented by parentheses embed grammars that contain syntactic restrictions; for example, the gender and number concordance typical of Spanish and which is preserved in nominal ellipsis [1]. Thus, the variable PRENOM (see Fig. 2) details the restrictions in the categories of the determiner and its concordance with the adjective.

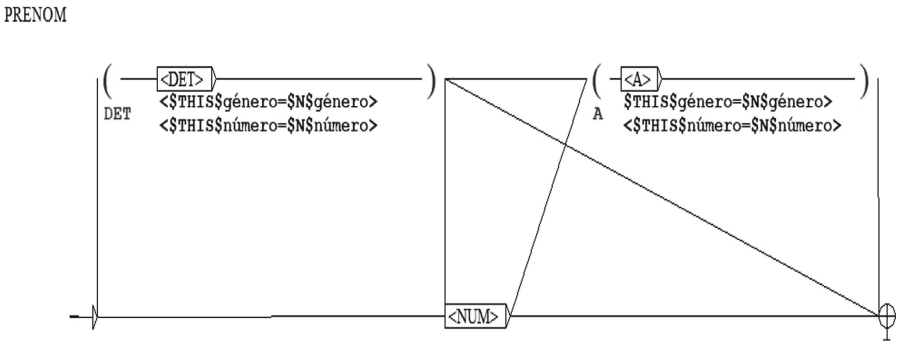


Fig. 2. Grammar embedded for determiner restrictions.

The POSTN variable indicates the restriction of the adjective with regards to its preceding N (see Fig. 3).

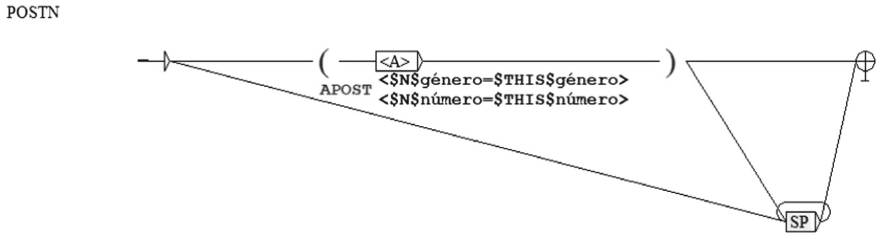


Fig. 3. Grammar embedded for adjective restrictions.

Finally, some grammars can be recursive (see Fig. 4). In the SP, the variation in the prepositional phrases found in the corpus can be described.

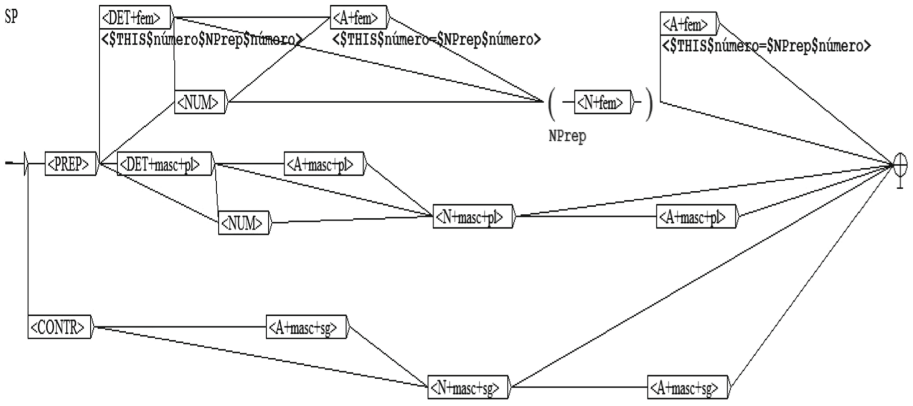


Fig. 4. Recursive grammars for phrase variation.

The indication of replacement implies a rewriting of the variables $PRENOM $NOM [1, 2, or 3] ($PREP) $DET; and later, of $NOM, although adjusted to the variables of number $DET. In cases in which a [1] $CONTR sequence appears, $NOM is replaced in the singular.

With the application of the grammar, it is possible to annotate (TAS) a grammatical sequence formally described with a corpus sequence (see Fig. 5).

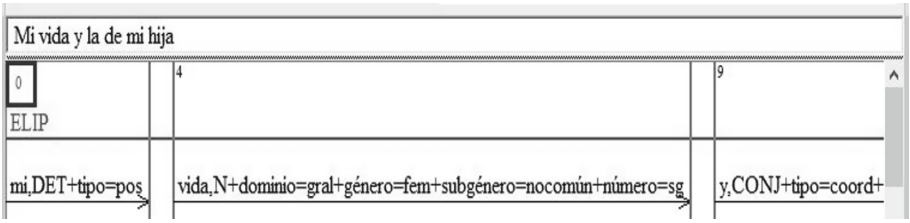


Fig. 5. Annotation of a grammatical sequence of the corpus

Furthermore, the implemented grammar recognizes, identifies, and replaces the elided element in the identified grammatical structure (see Fig. 6).

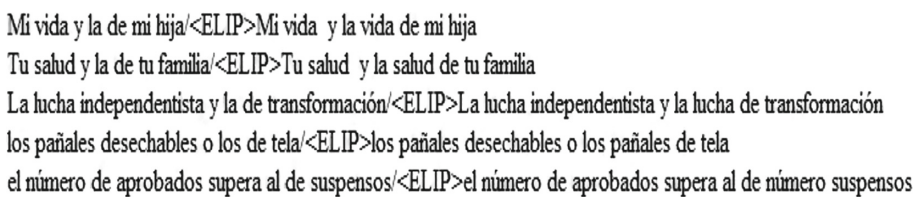


Fig. 6. Results of the automatic recognition of nominal ellipses

4 Results

Table 1 summarizes the main results of the computational implementation based on the collected texts, a brief corpus of 5,000 words with 100 elided sentences.

Table 1. Results obtained

Unidentified sentences	9
Mistakes	0
Coverage	82%
Precision	100%
F-measure	90.10%

As can be observed, the algorithm did not result in any incorrect labeling. Moreover, while there were coverage problems caused by lexical units not listed in the electronic dictionary, the results obtained show that the syntactic restrictions under which the nominal ellipsis is structured are useful for automatic identification in natural language texts. We also note that, in addition to the applications for syntactic recognition to optimize results in computational linguistics [26], it is also a tool that, beyond numerical data, provides information on grammatical knowledge.

5 Closing Remarks

This paper presented a computational modeling of nominal ellipsis in Spanish, based on a proposal for generative grammar [2], using an electronic dictionary and morphological grammars. The resulting algorithm showed a high percentage of precision, coverage, and F-measure. This implies that the descriptions stemming from formal studies constitute an adequate basis for the elaboration of computational devices. Despite these initially

promising results, the brevity of the corpus presents a limitation that merits further research.

Furthermore, in seeking to improve the ability of such models to resolve complex grammar when analyzing ellipses, future work will include comparative structures, such as (16), and those with preceding elements located in peripheral positions of the sentence, such as (17):

(15) *La gramática generativa es más compleja que la ~~gramática~~ de corte funcionalista*
[generative grammar is more complex than functionalist ~~grammar~~.]

(16) *Según los jugadores de Boca, los ~~jugadores~~ de River festejaron desmedidamente*
[according to Boca's players, River's ~~players~~ celebrated unreasonably.]

Lastly, we expect to include nouns in multi-word structures, as in (18).

(17) *La tasa de mortalidad de Chile es más alta que la ~~tasa de mortalidad~~ de Argentina*
[Chile's mortality rate is higher than Argentina's ~~mortality rate~~.]

For the latter, the complexity of the electronic dictionary will be increased with multi-word expressions that can also be linked to syntactic grammars.

Acknowledgments. This research was supported by a grant from the Proyecto Fondecyt Regular 1171033, from the Comisión Nacional de Investigación Científica y Tecnológica (Conicyt), Chile.

References

1. Saab, A.: Hacía una teoría de la identidad en la elipsis. Tesis de Doctorado. Universidad de Buenos Aires (2008)
2. Saab, A.: Nominal ellipsis. In: Craenenbroeck, J., Temmerman, T. (eds.) The Oxford Handbook of Ellipsis, pp. 526–561. Oxford University Press, Oxford (2019)
3. Silberztein, M.: Formalizing Natural Languages: The NooJ Approach. Wiley-Iste (2016)
4. Silberztein, M.: NooJ: a linguistic annotation system for corpus processing. Demo. In: Proceedings of the HLT/EMNLP2005 Conference, Vancouver (2005)
5. Merchant, J.: The syntax of silence: sluicing, Islands and identity in ellipsis. Tesis de Doctorado, Universidad de Santa Cruz (1999)
6. Sag, I.A., Hankamer, J.: Toward a theory of anaphoric processing. Linguist. Philos. **7**(3), 325–345 (1984). https://doi.org/10.1007/BF00627709
7. Culicover, P., Jackendoff, R.: Ellipsis in simpler syntax. In: Craenenbroeck, J., Temmerman, T. (eds.) The Oxford Handbook of Ellipsis, pp. 172–187. Oxford University Press, Oxford (2019)
8. Brucart, J.: La elipsis. In: Bosque, I., Demonte, V. (eds.) Gramática Descriptiva de la lengua español, vol. 1, no. 43, pp. 2787-2863. Espasa-Calpe, Madrid (1999)
9. Merchant, J.: The Syntax of SILENCE: SLUICING, ISLANDS, and the Theory of Ellipsis. Oxford University Press, Oxford (2001)
10. Chomsky, N.: Syntactic Structure. Walter de Gruyter, Berlin (1957)

11. Chomsky, N.: Aspects of the Theory of Syntax. MIT Press, Cambridge (1965)
12. Chomsky, N.: Lectures on Government and Binding. Gruyter Mouton, Berlin (1981)
13. Chomsky, N.: The Minimalist Program. MIT Press, Cambridge (1995)
14. Lobeck, A.: Ellipsis: Functional Heads, Licensing and Identification. Oxford University Press, New York (1995)
15. Ginzburg, J., Miller, P.: Ellipsis in head-driven phrase structure grammar. In: Craenenbroeck, J., Temmerman, T. (eds.) The Oxford Handbook of Ellipsis, pp. 75–121. Oxford University Press, Oxford (2019)
16. Culicover, P., Jackendoff, R.: Simpler Syntax. Oxford University Press, Oxford (2005)
17. Hardt, D.: Verb phrase ellipsis: form, meaning, and processing. Tesis de Doctorado, Universidad de Pennsylvania (1993)
18. Mitkov, R.: The Oxford Handbook of Computational Linguistics. Oxford University Press, Oxford (2002)
19. Nielsen, L.: Verb phrase ellipsis detection using automatically parsed text. In: COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics, Geneva, pp.1093–1099 (2004)
20. Rello, L., Baeza-Yates, R., Mitkov, R.: Elliphant: improved automatic detection of zero subjects and impersonal constructions in Spanish. In: Conference: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp. 706-715 (2012)
21. McShane, M., Babkin, P.: Detection and resolution of verb phrase ellipsis. LiLT **13**(1), 1–34 (2016)
22. Merchant, J.: Ellipsis: a survey of analytical approaches. In: Craenenbroeck, J., Temmerman, T. (eds.) Handbook of Ellipsis. Oxford University Press, Oxford (2016)
23. Halle, M., Marantz, A.: Distributed morphology and pieces of inflection. In: Hale, K., Keyser, S. (eds.) The View from Building 20, pp. 111–176. MIT Press, Cambridge (1993)
24. Abney, S.: The English noun phrase in its sentential aspect. Tesis doctoral, MIT (1987)
25. Real Academia Española: Diccionario de la Real academia de la lengua Española. Espasa, Madrid (2014)
26. Hardt, D.: Ellipsis and computational linguistics. In: Craenenbroeck, J., Temmerman, T. (eds.) The Oxford Handbook of Ellipsis, pp. 342–356. Oxford University Press, Oxford (2019)