

# A Computational Approach to Zero-pronouns in Spanish

Antonio Ferrández and Jesús Peral

Dept. Languages and Information Systems, University of Alicante

Carretera San Vicente S/N

03080 ALICANTE, Spain

{antonio, jperal}@dlsi.ua.es

## Abstract

In this paper, a computational approach for resolving zero-pronouns in Spanish texts is proposed. Our approach has been evaluated with partial parsing of the text and the results obtained show that these pronouns can be resolved using similar techniques that those used for pronominal anaphora. Compared to other well-known baselines on pronominal anaphora resolution, the results obtained with our approach have been consistently better than the rest.

## Introduction

In this paper, we focus specifically on the resolution of a linguistic problem for Spanish texts, from the computational point of view: *zero-pronouns* in the “subject” grammatical position. Therefore, the aim of this paper is not to present a new theory regarding zero-pronouns, but to show that other algorithms, which have been previously applied to the computational resolution of other kinds of pronoun, can also be applied to resolve zero-pronouns.

The resolution of these pronouns is implemented in the computational system called *Slot Unification Parser for Anaphora resolution (SUPAR)*. This system, which was presented in Ferrández et al. (1999), resolves anaphora in both English and Spanish texts. It is a modular system and currently it is being used for Machine Translation and Question Answering, in which this kind of pronoun is very important to solve due to its high frequency in Spanish texts as this paper will show.

We are focussing on zero-pronouns in Spanish texts, although they also appear in other languages, such as Japanese, Italian and Chinese. In English texts, this sort of pronoun occurs far less frequently, as the use of subject

pronouns is generally compulsory in the language. While in other languages, zero-pronouns may appear in either the subject’s or the object’s grammatical position, (e.g. Japanese), in Spanish texts, zero-pronouns only appear in the position of the subject.

In the following section, we present a summary of the present state-of-the-art for zero-pronouns resolution. This is followed by a description of the process for the detection and resolution of zero-pronouns. Finally, we present the results we have obtained with our approach.

## 1 Background

Zero-pronouns have already been studied in other languages, such as Japanese, (e.g. Nakaiwa and Shirai (1996)). They have not yet been studied in Spanish texts, however. Among the work done for their resolution in different languages, nevertheless, there are several points that are common for Spanish. The first point is that they must first be located in the text, and then resolved. Another common point among, they all employ different kinds of knowledge (e.g. morphologic or syntactic) for their resolution. Some of these works are based on the Centering Theory (e.g. Okumura and Tamura (1996)). Other works, however, distinguish between *restrictions* and *preferences* (e.g. Lappin and Leass (1994)). *Restrictions* tend to be absolute and, therefore, discard any possible antecedents, whereas *preferences* tend to be relative and require the use of additional criteria, i.e. heuristics that are not always satisfied by all anaphors. Our anaphora resolution approach belongs to the second group.

In computational processing, semantic and domain information is computationally inefficient when compared to other kinds of knowledge. Consequently, current anaphora resolution methods rely mainly on restrictions and preference heuristics, which employ

information originating from morpho-syntactic or shallow semantic analysis, (see Mitkov (1998) for example). Such approaches, nevertheless, perform notably well. Lappin and Leass (1994) describe an algorithm for pronominal anaphora resolution that achieves a high rate of correct analyses (85%). Their approach, however, operates almost exclusively on syntactic information. More recently, Kennedy and Boguraev (1996) propose an algorithm for anaphora resolution that is actually a modified and extended version of the one developed by Lappin and Leass (1994). It works from a POS tagger output and achieves an accuracy rate of 75%.

## 2 Detecting zero-pronouns

In order to detect zero-pronouns, the sentences should be divided into clauses since the subject could only appear between the clause constituents. After that, a noun-phrase (NP) or a pronoun that agrees in person and number with the clause verb is sought, unless the verb is imperative or impersonal.

As we are also working on unrestricted texts to which partial parsing is applied, zero-pronouns must also be detected when we do not dispose of full syntactic information. In Ferrández et al. (1998), a partial parsing strategy that provides all the necessary information for resolving anaphora is presented. That study shows that only the following constituents were necessary for anaphora resolution: co-ordinated prepositional and noun phrases, pronouns, conjunctions and verbs, regardless of the order in which they appear in the text.

**H1** Let us assume that the beginning of a new clause has been found when a verb is parsed and a free conjunction is subsequently parsed.

When partial parsing is carried out, one problem that arises is to detect the different clauses of a sentence. Another problem is how to detect the zero-pronoun, i.e. the omission of the subject from each clause. With regard to the first problem, the heuristic *H1* is applied to identify a new clause.

(1) John y Jane llegaron tarde al trabajo **porque**  $\emptyset^1$  se durmieron (John and Jane were late for work **because** [they] $\emptyset$  over-slept)

<sup>1</sup> The symbol  $\emptyset$  will always show the position of the

In this particular case, a *free conjunction* does not imply conjunctions<sup>2</sup> that join co-ordinated noun and prepositional phrases. It refers, here, to conjunctions that are parsed in our partial parsing scheme. For instance, in sentence (1), the following sequence of constituents is parsed:

*np(John and Jane), verb(were), freeWord<sup>3</sup>(late), pp(for work), conj(because), pron(they), verb(over-slept)*

Since the free conjunction *porque* (*because*) has been parsed after the verb *llegaron* (*were*), the new clause with a new verb *durmieron* (*over-slept*) can be detected.

With reference to the problem about detecting the omission of the subject from each clause with partial parsing, it is solved by searching through the clause constituents that appear before the verb. In sentence (1), we can verify that the first verb, *llegaron* (*were*), does not have its subject omitted since there appears a *np(John and Jane)*. However, there is a zero-pronoun, *(they) $\emptyset$* , for the second verb *durmieron* (*over-slept*).

(2) Pedro<sub>j</sub> vio a Ana<sub>k</sub> en el parque.  $\emptyset_k$  Estaba muy guapa (Peter<sub>j</sub> saw Ann<sub>k</sub> in the park. [She] <sub>$\emptyset_k$</sub>  was very beautiful)

When the zero-pronoun is detected, our computational system inserts the pronoun in the position in which it has been omitted. This pronoun will be resolved in the following module of anaphora resolution. Person and number information is obtained from the clause verb. Sometimes in Spanish, gender information of the pronoun can be obtained when the verb is copulative. For example, in sentence (2), the verb *estaba* (*was*) is copulative, so that its subject must agree in gender and number with its object whenever the object can have either a masculine or a feminine linguistic form (*guapo*: *masc*, *guapa*: *fem*). We can therefore obtain information about its gender from the object, *guapa* (*beautiful* in its feminine form) which automatically assigns it to the feminine gender so the omitted pronoun would have to be *she* rather than *he*. Gender information can be obtained from the object of the verb with partial

omitted pronoun.

<sup>2</sup> For example, it would include punctuation marks such as a semicolon.

<sup>3</sup> The *free words* consist of constituents that are not covered by this partial parsing (e.g. adverbs).

parsing as we simply have to search for a NP on the right of the verb.

### 3 Zero-pronoun resolution

In this module, *anaphors* (i.e. anaphoric expressions such as pronominal references or zero-pronouns) are treated from left to right as they appear in the sentence, since, at the detection of any kind of anaphor, the appropriate set of restrictions and preferences begins to run.

The number of previous sentences considered in the resolution of an anaphora is determined by the kind of anaphora itself. This feature was arrived at following an in depth study of Spanish texts. For pronouns and zero-pronouns, the antecedents in the four previous sentences, are considered.

The following restrictions are first applied to the list of candidates: person and number agreement, c-command<sup>4</sup> constraints and semantic consistency<sup>5</sup>. This list is sorted by proximity to the anaphor. Next, if after applying the restrictions there is still more than one candidate, the preferences are then applied, with the degree of importance shown in Figure 1.

This sequence of preferences (from 1 to 10) stops whenever only one candidate remains after having applied a given preference. If after all the preferences have been applied there is still more than one candidate left, the most repeated candidates<sup>6</sup> in the text are then extracted from the list, and if there is still more than one candidate, then the candidates that have appeared most frequently with the verb of the anaphor are extracted from the previous list. Finally, if after having applied all the previous preferences, there is still more than one candidate left, the first candidate of the resulting list (the closest to the anaphor) is selected.

The set of constraints and preferences required for Spanish pronominal anaphora presents two basic differences: a) zero-pronoun resolution has the restriction of agreement only

in person and number, (whereas pronominal anaphora resolution requires gender agreement as well), and b) a different set of preferences.

- 1) Candidates in the same sentence as the anaphor.
- 2) Candidates in the previous sentence.
- 3) Preference for candidates in the same sentence as the anaphor and those that have been the solution of a zero-pronoun in the same sentence as the anaphor.
- 4) Preference for proper nouns or indefinite NPs.
- 5) Preference for proper nouns.
- 6) Candidates that have been repeated more than once in the text.
- 7) Candidates that have appeared with the verb of the anaphor more than once.
- 8) Preference for noun phrases that are not included in a prepositional phrase or those that are connected to an Indirect Object.
- 9) Candidates in the same position as the anaphor, with reference to the verb (before the verb).
- 10) If the zero-pronoun has gender information, those candidates that agree in gender.

Figure 1. Anaphora resolution preferences.

The main difference between the two sets of preferences is the use of two new preferences in our algorithm: Nos. 3 and 10. Preference 10 is the last preference since the POS tagger does not indicate whether the object has both masculine and feminine linguistic forms<sup>7</sup> (i.e. information obtained from the object when the verb is copulative). Gender information must therefore be considered a preference rather than a restriction. Another interesting fact is that syntactic parallelism (Preference No. 9) continues to be one of the last preferences, which emphasizes the unique problem that arises in Spanish texts, in which syntactic structure is quite flexible (unlike English).

## 4 Evaluation

### 4.1 Experiments accomplished

Our computational system (SUPAR) has been trained with a handmade corpus<sup>8</sup> with 106 zero-

<sup>4</sup> The usage of c-command restrictions on partial parsing is presented in Ferrández et. al. (1998).

<sup>5</sup> Semantic knowledge is only used when working on restricted texts.

<sup>6</sup> Here, we mean that we first obtain the maximum number of repetitions for an antecedent in the remaining list. After that, we extract the antecedents that have this value of repetition from the list.

<sup>7</sup> For example in: *Peter es un genio* (*Peter is a genius*), the tagger does not indicate that the object does not have both masculine and feminine linguistic forms. Therefore, a feminine subject would use the same form: *Jane es un genio* (*Jane is a genius*). Consequently, although the tagger says that the verb, *es* (*is*), is copulative, and the object, *un genio* (*a genius*) is masculine, this gender could not be used as a restriction for the zero-pronoun in the following sentence:  $\emptyset$  *Es un genio*.

<sup>8</sup> This corpus has been provided by our colleagues in

pronouns. This training has mainly supposed the improvement of the set of preferences, i.e. the optimum order of preferences in order to obtain the best results. After that, we have carried out a blind evaluation on unrestricted texts. Specifically, SUPAR has been run on two different Spanish corpora: a) a part of the Spanish version of *The Blue Book* corpus, which contains the handbook of the International Telecommunications Union CCITT, published in English, French and Spanish, and automatically tagged by the Xerox tagger, and b) a part of the *Lexesp* corpus, which contains Spanish texts from different genres and authors. These texts are taken mainly from newspapers, and are automatically tagged by a different tagger than that of *The Blue Book*. The part of the *Lexesp* corpus that we processed contains ten different stories related by a sole narrator, although they were written by different authors. Having worked with different genres and disparate authors, we feel that the applicability of our proposal to other sorts of texts is assured. In Figure 2, a brief description of these corpora is given. In these corpora, partial parsing of the text with no semantic information has been used.

	Number of words	Number of sentences	Words per sentence
Lexesp corpus Text 1	972	38	25.6
Text 2	999	55	18.2
Text 3	935	34	27.5
Text 4	994	36	27.6
Text 5	940	67	14
Text 6	957	34	28.1
Text 7	1025	59	17.4
Text 8	981	40	24.5
Text 9	961	36	26.7
Text 10	982	32	30.7
The Blue Book corpus	15,571	509	30.6

Figure 2. Description of the unrestricted corpora used in the evaluation.

## 4.2 Evaluating the detection of zero-pronouns

To achieve this sort of evaluation, several different tasks may be considered. Each verb must first be detected. This task is easily

the University of Alicante, which were required to propose sentences with zero-pronouns.

accomplished since both corpora have been previously tagged and manually reviewed. No errors are therefore expected on verb detection. Therefore, a *recall*<sup>9</sup> rate of 100% is accomplished. The second task is to classify the verbs into two categories: a) verbs whose subjects have been omitted, and b) verbs whose subjects have not. The overall results on this sort of detection are presented in Figure 3 (*success*<sup>10</sup> rate of 88% on 1,599 classified verbs, with no significant differences seen between the corpora). We should also remark that a success rate of 98% has been obtained in the detection of verbs whose subjects were omitted, whereas only 80% was achieved for verbs whose subjects were not. This lower success rate is justified, however, for several reasons. One important reason is the non-detection of impersonal verbs by the POS tagger. This problem has been partly resolved by heuristics such as a set of impersonal verbs (e.g. *llover* (to rain)), but it has failed in some impersonal uses of some verbs. For example, in sentence (3), the verb *es* (to be) is not usually impersonal, but it is in the following sentence, in which SUPAR would fail:

(3)  $\emptyset$  Es hora de desayunar ([It] $\emptyset$  is time to have breakfast)

Two other reasons for the low success rate achieved with verbs whose subjects were not omitted are the lack of semantic information and the inaccuracy of the grammar used. The second reason is the ambiguity and the unavoidable incompleteness of the grammars, which also affects the process of clause splitting.

In Figure 3, an interesting fact can be observed: 46% of the verbs in these corpora have their subjects omitted. It shows quite clearly the importance of this phenomenon in Spanish. Furthermore, it is even more important in narrative texts, as this figure shows: 61% with the Lexesp corpus, compared to 26% with the technical manual. We should also observe that The Blue Book has no verbs in either the first or the second person. This may be explained by the style of the technical manual, which usually

<sup>9</sup> By “recall rate”, we mean the number of verbs classified, divided by the total number of verbs in the text.

<sup>10</sup> By “success rate”, we mean the number of verbs successfully classified, divided by the total number of verbs in the text.

	Verbs with their subject omitted						Verbs with their subject <b>no</b> -omitted					
	First person		Second person		Third person		First person		Second person		Third person	
	Total	% Success	Total	% Success	Total	% Success	Total	% Success	Total	% Success	Total	% Success
Lexesp corpus	111	100%	42	100%	401	99%	21	81%	3	100%	328	76%
	20%		7%		73%		7%		1%		92%	
	554 (61%) (success rate: 99%)						352 (39%) (success rate: 76%)					
Blue Book corpus	0	0%	0	0%	180	97%	0	0%	0	0%	513	82%
	0%		0%		100%		0%		0%		100%	
	180 (26%) (success rate: 97%)						513 (74%) (success rate: 82%)					
Total	734 (46%) (success rate: 98%)						865 (54%) (success rate: 80%)					
	1,599 (success rate: 88%)											

Figure 3. Results obtained in the detection of zero-pronouns.

consists of a series of isolated definitions, (i.e. many paragraphs that are not related to one another). This explanation is confirmed by the relatively small number of anaphors that are found in that corpus, as compared to the Lexesp corpus.

We have not considered comparing our results with those of other published works, since, (as we have already explained in the *Background* section), ours is the first study that has been done specifically for Spanish texts, and the designing of the detection stage depends mainly on the structure of the language in question. Any comparisons that might be made concerning other languages, therefore, would prove to be rather insignificant.

### 4.3 Evaluating anaphora resolution

As we have already shown in the previous section, (Figure 3), of the 1,599 verbs classified in these two corpora, 734 of them have zero-pronouns. Only 581 of them, however, are in third person and will be resolved. In Figure 4, we present a classification of these third person zero-pronouns, which have been conveniently divided into three categories: cataphoric, exophoric and anaphoric. The first category is comprised of those whose antecedent, i.e. the clause subject, comes after the verb. For example, in sentence (4) the subject, *a boy*, appears after the verb *compró* (*bought*).

(4)  $\emptyset_k$  Compró un niño<sub>k</sub> en el supermercado (A boy<sub>k</sub> bought in the supermarket)

This kind of verb is quite common in Spanish, as can be seen in this figure (49%). This fact represents one of the main difficulties found in resolving anaphora in Spanish: the structure of a sentence is more flexible than in English. These represent intonationally marked sentences, where the subject does not occupy its usual position in the sentence, i.e. before the verb. Cataphoric zero-pronouns will not be resolved in this paper, since semantic information is needed to be able to discard all of their antecedents and *to prefer* those that appear within the same sentence and clause after the verb. For example, sentence (5) has the same syntactic structure than sentence (4), i.e. *verb, np, pp*, where the subject function of the *np* can only be distinguished from the object by means of semantic knowledge.

(5)  $\emptyset$  Compró un regalo en el supermercado ([He] <sub>$\emptyset$</sub>  bought a present in the supermarket)

The second category consists of those zero-pronouns whose antecedents do not appear, linguistically, in the text (they refer to items in the external world rather than things referred to in the text). Finally, the third category is that of pronouns that will be resolved by our computational system, i.e., those whose antecedents come before the verb: 228 zero-pronouns. These pronouns would be equivalent to the full pronoun *he, she, it or they*.

	Cataphoric	Exophoric	Anaphoric	
			Number	Success
Lexesp corpus	171 (42%)	56 (12%)	174 (46%)	78%
The Blue Book corpus	113 (63%)	13 (7%)	54 (30%)	68%
Total	284 (49%)	69 (12%)	228 (39%)	75%

Figure 4. Classification of third person zero-pronouns.

The different accuracy results are also shown in Figure 4: A success rate of 75% was attained for the 228 zero-pronouns. By “successful resolutions” we mean that the solutions offered by our system agree with the solutions offered by two human experts.

For each zero-pronoun there is, on average, 355 candidates before the restrictions are applied, and 11 candidates after restrictions. Furthermore, we repeated the experiment without applying restrictions and the success rate was significantly reduced.

Since the results provided by other works have been obtained on different languages, texts and sorts of knowledge (e.g. Hobbs and Lappin full parse the text), direct comparisons are not possible. Therefore, in order to accomplish this comparison, we have implemented some of these approaches in SUPAR. Although some of these approaches were not proposed for zero-pronouns, we have implemented them since as our approach they could also be applied to solve this kind of pronoun. For example, with the baseline presented by Hobbs (1977) an accuracy of 49.1% was obtained, whereas, with our system, we achieved 75% accuracy. These results highlight the improvement accomplished with our approach, since Hobbs’ baseline is frequently used to compare most of the work done on anaphora resolution<sup>11</sup>. The reason why Hobbs’ algorithm works worse than ours is due to the fact that it carries out a full parsing of the text. Furthermore, the way to explore the syntactic tree with Hobbs’ algorithm is not the best one for the Spanish language since it is nearly a free-word-order language.

Our proposal has also been compared with the typical baseline of morphological agreement and proximity preference, (i.e., the antecedent

that appears closest to the anaphor is chosen from among those that satisfy the restrictions). The result is a 48.6% accuracy rate. Our system, therefore, improves on this baseline as well.

Lappin and Leass (1994) has also been implemented in our system and an accuracy of 64% was attained. Moreover, in order to compare our proposal with Centering approach, Functional Centering by Strube and Hahn (1999) has also been implemented, and an accuracy of 60% was attained.

One of the improvements afforded by our proposal is that statistical information from the text is included with the rest of information (syntactic, morphologic, etc.). Dagan and Itai (1990), for example, developed a statistical approach for pronominal anaphora, but the information they used was simply the patterns obtained from the previous analysis of the text. To be able to compare our approach to that of Dagan and Itai, and to be able to evaluate the importance of this kind of information, our method was applied with statistical information<sup>12</sup> only. If there is more than one candidate after applying statistical information, *preference*, and then *proximity preference* are applied. The results obtained were lower than when all the preferences are applied jointly: 50.8%. These low results are due to the fact that statistical information has been obtained from the beginning of the text to the pronoun. A previous training with other texts would be necessary to obtain better results.

Regarding the success rates reported in Ferrández et al. (1999) for pronominal references (82.2% for Lexesp, 84% for Spanish version of The Blue Book, and 87.3% for the English version), are higher than our 75% success rate for zero-pronouns. This reduction (from 84% to 75%) is due mainly to the lack of gender information in zero-pronouns.

Mitkov (1998) obtains a success rate of 89.7% for pronominal references, working with English technical manuals. It should be pointed out, however, that he used some knowledge that was very close to the genre<sup>13</sup> of the text. In our

<sup>11</sup> In Tetreault (1999), for example, it is compared with an adaptation of the Centering Theory by Grosz et al. (1995), and Hobbs’ baseline out-performs it.

<sup>12</sup> This statistical information consists of the number of times that a word appears in the text and the number of times that it appears with a verb.

<sup>13</sup> For example, the antecedent indicator *section heading preference*, in which if a NP occurs in the heading of the section, part of which is the current

study, such information was not used, so we consider our approach to be more easily adaptable to different kinds of texts. Moreover, Mitkov worked exclusively with technical manuals whereas we have worked with narrative texts as well. The difference observed is due mainly to the greater difficulty found in narrative texts than in technical manuals which are generally better written. In any case, the applicability of our proposal to different genres of texts seems to have been well proven. Anyway, if the order of application of the preferences<sup>14</sup> is varied to each different text, an 80% overall accuracy rate is attained. This fact implies that there is another kind of knowledge, close to the genre and author of the text that should be used for anaphora resolution.

## Conclusion

In this paper, we have proposed the first algorithm for the resolution of zero-pronouns in Spanish texts. It has been incorporated into a computational system (SUPAR). In the evaluation, several baselines on pronominal anaphora resolution have been implemented, and it has achieved better results than either of them have.

As a future project, the authors shall attempt to evaluate the importance of semantic information for zero-pronoun resolutions in unrestricted texts. Such information will be obtained from a lexical tool, (e.g. EuroWordNet), which could be consulted automatically. We shall also evaluate our proposal in a Machine Translation application, where we shall test its success rate by its generation of the zero-pronoun in the target language, using the algorithm described in Peral et al. (1999).

## References

Ido Dagan and Alon Itai (1990) Automatic processing of large corpora for the resolution of anaphora references. In *Proceedings of the 13<sup>th</sup>*

*International Conference on Computational Linguistics, COLING* (Helsinki, Finland).

Antonio Ferrández, Manuel Palomar and Lidia Moreno (1998) Anaphora resolution in unrestricted texts with partial parsing. In *Proceedings of the 36<sup>th</sup> Annual Meeting of the Association for Computational Linguistics and 17<sup>th</sup> International Conference on Computational Linguistics, COLING - ACL* (Montreal, Canada). pp. 385-391.

Antonio Ferrández, Manuel Palomar and Lidia Moreno (1999) An empirical approach to Spanish anaphora resolution. *To appear in Machine Translation 14(2-3)*.

Jerry Hobbs (1977) Resolving pronoun references. *Lingua*, 44. pp. 311-338.

Cristopher Kennedy and Bran Boguraev (1996) Anaphora for Everyone: Pronominal Anaphora resolution without a Parser. In *Proceedings of the 16<sup>th</sup> International Conference on Computational Linguistics, COLING* (Copenhagen, Denmark). pp. 113-118.

Shalom Lappin and Herb Leass (1994) An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4). pp. 535-561.

Ruslan Mitkov (1998) Robust pronoun resolution with limited knowledge. In *Proceedings of the 36<sup>th</sup> Annual Meeting of the Association for Computational Linguistics and 17<sup>th</sup> International Conference on Computational Linguistics, COLING - ACL* (Montreal, Canada). pp. 869-875.

Hiromi Nakaiwa and Satoshi Shirai (1996) Anaphora Resolution of Japanese Zero Pronouns with Deictic Reference. In *Proceedings of the 16<sup>th</sup> International Conference on Computational Linguistics, COLING* (Copenhagen, Denmark). pp. 812-817.

Manabu Okumura and Kouji Tamura (1996) Zero Pronoun Resolution in Japanese Discourse Based on Centering Theory. In *Proceedings of the 16<sup>th</sup> International Conference on Computational Linguistics, COLING* (Copenhagen, Denmark). pp. 871-876.

Jesús Peral, Manuel Palomar and Antonio Ferrández (1999) Coreference-oriented Interlingual Slot Structure and Machine Translation. In *Proceedings of ACL Workshop on Coreference and its Applications* (College Park, Maryland, USA). pp. 69-76.

Michael Strube and Udo Hahn (1999) Functional Centering – Grounding Referential Coherence in Information Structure. *Computational Linguistics*, 25(5). pp. 309-344.

---

sentence, it is considered to be the preferred candidate.

<sup>14</sup> The difference between the individual sets of preferences is the degree of importance of the preferences for proper nouns and syntactic parallelism.