ELLIPSIS IN SPONTANEOUS SPOKEN LANGUAGE

Manuel Alcántara*& Núria Bertomeu
Universidad Autónoma de Madrid; Universität des Saarlandes
manuel@maria.lllf.uam.es; bertomeu@coli.uni-sb.de

Abstract

In this paper we present an empirical study of elliptical phenomena carried out on a spontaneous speech corpus of Spanish annotated with event structures and co-references. The latter have allowed us to automatically identify elliptical fragments, as well as their antecedents and to measure the distance in sentences and turns between them. We also have looked into the nature of elliptical utterances without an explicit linguistic source and to their relation with the context.

1 Introduction

Ellipsis is an implicit reference to some material, either previously mentioned, or somehow inferable. In both cases, this elided material can be successfully recovered and applied to the remnant of the ellipsis, that is, to the piece of information stated explicitly in the elliptical fragment.

Regarding the first case, it is necessary for the elided material to be *accessible* in the dialogue participant's discourse record, that is, activated in the focus of attention or working memory. Some models predict that accessibility of the source is given by the discourse-structure. In (Schlangen 2003) if an utterance is a possible attachment point upon the *right frontier constraint* it can behave as a source. Hardt and Romero (2004) claim that antecedents must c-command elliptical clauses in the discourse-tree. In (Cooper and Ginzburg 2001) the antecedent of an elliptical fragment must be the maximal question under discussion (MAX-QUD), and when this is not like this, resolving a fragment amounts to accommodating the right MAX-QUD (Cooper et al. 2000).

On the other hand, recent experimental results from the neuro-psycholinguistic side (Streb et al. 2004) show that ellipsis resolution involves syntactic processing, unlike anaphora resolution, which involves semantic processing. This means that the syntactic representation of the source must be still in working-memory at the time when the elliptical fragment is uttered. The syntactic structure of a sentence usually remains in memory for a very short time, a little bit longer than the phonological form, and then decays, unlike the semantic representation of the sentence, which remains longer in memory (Kintsch and van Dijk 1978). However, it is not clear how long syntactic structure keeps being active.

Some of the questions we attempt to answer with this corpus study are how long the distance between source and ellipsis can be, which is the most frequent distance, and

*The author wants to express his acknowledgement to the RILARIM project (TIN2004-07588-C03-02) for partially supporting his research.

which is the nature of the accessibiliy of the source - whether the presence of its syntactic structure in memory, or it being an attachment point in the discourse structure, or whether both levels of representation complement each other.

We also will consider cases of ellipsis without an antecedent in the preceding discourse, where the missing part has to be inferred, and look at their relation to the context, i.e. whether they refer to some previously mentioned entity, as well as find out which are the most frequently omitted lemmas.

Finally, we want to find out whether remnants tend to be of a particular argument type. Following Centering Theory (Grosz et al. 1995), less oblique argument types carry referential continuity, while entities referred to in more oblique arguments tend to be discourse and/or hearer new, and to be expressed explicitly. Besides, remnants usually convey new information (in the information-structure sense). This new information is not necessarily discourse/hearer new, but in many cases it is. Upon this, we would expect remnants of ellipsis to be more oblique arguments types.

In section 2, we speak about the corpus and the methodology we have employed to carry out our experiments. In section 3 we present the results for the three topics presented above and in section 4 we discuss them and conclude.

2 Corpus and Methodology

The present experiments have been carried out on a 50000 words subcorpus of the Spanish part of C-ORAL-ROM (Moreno et al. 2005), a corpus of spoken language which was recorded following strict requirements of spontaneity and variety of speakers and contexts. This corpus represents a wide variety of speech acts performed in the daily use of language.

Sentence tokenization has been done following semantic constraints and every sentence corresponds to a complete event-structure[1]. That's why a sentence may contain several turns if these are overlapping and interrupt its utterance. It is also distinguished between the continuation of a sentence after a filled pause and a fragment which elaborates on a previous sentence according to pause lengths. The transcriptions are manually annotated with semantic information following the tag set of SESCO, a tagging system which allows the semantic representation of linguistic corpora (Alcántara 2005).

SESCO provides general information about predicates and their argument structures. This tagging system follows a compositional approach based on event structures. Events are classified under only three major types: states, processes and actions, and these major types can be divided into subtypes according to the arguments they require. This approach is compositional because a state has two arguments (an entity and its property/location), a process is made up of a transition from one state to another, and an action is a process with an agent and a patient; besides, those parts of an event which are not arguments are tagged as indirect relations.

SESCO has particular tags for linking different references related to the same entity or event. The first time an element is mentioned an identifier (IDE) is assigned to it by means of a unique code. The following ocurrences of this element are annotated with a

[1] In the rest of the paper we will refer as sentences to those semantically complete main event structures.

reference (REF) to the same code. We have used these tags in order to identify elliptical utterances and their sources and to study the distances between them.

In our study we have considered cases of intra-sentential and inter-sentential ellipsis. We have looked at cases of verbal ellipsis[2]. The main strategy was to find those events lacking a finite verb lexeme and being co-referent with another event. The second strategy was to find those events lacking a finite verb lexeme and not co-referent with any previous event, and look for entities within them being co-referent with some other entity recently mentioned in the discourse. Each strategy returned a different kind of fragment, those with an explicit linguistic source and those without it, respectively. All information was automatically retrieved and the results manually checked to ensure that the tagging was correct.

3 Experiments and results

From a total of 6922 events in the corpus 522 were found to be elliptical. This corresponds with a 7.5%, a considerable amount which confirms the importance of ellipsis resolution when attempting to understand spontaneous spoken texts. The absolute frequency of elliptical utterances with an antecedent is 306, and of those without, 216 - in relative terms, 58.62% and 41.38%, respectively.

3.1 Ellipsis with Antecedent

With the aim of finding out how far an elliptical construction can be from its antecedent we calculated the distance in sentences and turns from the ellipsis to the first and last preceding occurrences of the event co-referent with it. The idea of calculating the last preceding occurrence aroused from the hypothesis that fragments can behave themselves as sources, that is, that they keep accessible the structure implicitly expressed in them. We decided to calculate also the distance from the first occurrence of the event in order to study the information that these elliptical sources carry from their own source and the changes they introduce to it. Of course, when there was only one preceding occurrence of the event, only this one was considered. The distances from the last preceding occurrence of the event are the following:

Distance (sentences)	total	percentage
1	219	71.57%
2	43	14.05%
3	16	5.23%
0	14	4.58%
4	11	3.59%
5	1	0.33%
6	1	0.33%
16	1	0.33%

As shown in the table the most frequent distance is 1 sentence, followed by 2, 3, 0 (intra-sentential ellipsis) and 4. However, we also find one case of 16 sentences of distance.

[2]We leave fragments formed by the alone-standing affirmative and negative adverbs (yes/no) and nominal ellipsis - null-objects and semantic ellipsis - for future research.

According to what hierarchical models of discourse structure would predict (see for example (Schlangen 2003)) in this case the fragment and all the intervening material are subordinated to the source.

These results are not very different from the ones obtained when calculating the distance to the first occurrence of the event, since in 196 of 289 cases there is only one preceding occurrence. The maximum distance to the first occurrence is also 16 sentences and 6, 7, 8 and 9 sentences have frequencies of 1.7%, 2%, 1.03% and 1.03%, respectively. In most cases the following occurrences of the event are elliptical themselves. It was worth to look at them because of the interest of cases like the following:

(1) BLA: Quedamos en la escuela?
Meet at the school
'Shall we meet at school?'
YOL: O si te viene mejor en Moncloa?
Or if you suit better in Moncloa
'Or if it suits you better (to meet) in Moncloa (we can meet in Moncloa)'
BLA: Vale, mejor.
O.K. better.
'O.K., (we'd) better (meet in Moncloa).'
YOL: Como el otro día y así no tienes que subir ni nada.
Like the other day and so not have to go up neither nothing
'(We are meeting) Like the other day, and so you don't need to come upstairs.'
BLA: Vale. En como el viernes.
O.K. In like the Friday.
'O.K. (We are meeting) Where (we met) on Friday.'
YOL: Sí. Quedamos a menos diez o menos cuarto.
Yes. Meet minus ten or minus quarter.
'Yes. Let's meet at ten to or quarter to.'
BLA: A las ocho menos diez?
At the eight minus ten?
'(Are we meeting) At ten to eight?'

This example is interesting for two reasons. In the second fragment we find a conditional sentence which is itself elliptical. Moreover, the missing main sentence is structurally identical with the source, but it must be inferred that in the resolution 'Moncloa' substitutes 'en la escuela'. It doesn't matter how the following fragment is resolved, but it establishes the fact that the meeting is going to be in Moncloa[3]. So a full resolution of the last fragments has to contain that the meeting is taking place in Moncloa, although this is not explicitly said anywhere. That is, although there is structural identity, some kind of reasoning is needed in order to successfully resolve those fragments. This is an example of how successive material changes the original event and, thus, a full explicit source is not to be found.

[3]Here and in the examples 2 and 5, we have choosen to write between parenthesis what we think could be a surface resolution of the fragment, however it is not very clear, at least for the surface, how these fragments should be resolved, though they are semantically unambiguous.

We also calculated the distance in turns between source and fragments. We excluded monologues, since there is no turn-taking in them. Again we calculated the distance from the first and last occurrences of the event previous to the fragment. Figures for the former are not shown because they are very similar to those for the latter. For the latter we found distances of up to 7 turns with the most frequent being 1 turn (42.15%), followed by 0 (36.27%), 2 (11.76%), 3 (4.24%), 4 (2.65%), 5 (1.63%), 6 (0.65%) and 7 (0.65%).

In order to see whether the speaker's own utterances are more accessible as antecedents, we looked at the frequency and distance between those fragment-antecedent pairs uttered by the same speaker and those uttered by different speakers. For calculating the frequency we left monologues away again. In our corpus both kinds of pairs have aproximately the same frequency, 49.4% and 50.5% respectively. However, the range of distances for those uttered by the same speaker is a little bit larger than for those uttered by different speakers. For a single speaker we find distances up to 16 sentences, while for several speakers up to 9. However, we do not think this fact is sufficient to conclude anything since those distances up to 9 for a single speaker are quite unimportant in number. Moreover, while the frequencies for 2 sentences of distance are higher for one speaker, those of 3, 4 and 5 are higher for several speakers; 7 is then higher for one speaker and 8 and 9 for several. Our data do not show, thus, any clear tendency, but rather speak for a great degree of similarity among the discourse representations of the different dialogue participants.

3.2 Ellipsis without antecedent

For those cases of ellipsis without an explicit linguistic source we took as starting point the hypothesis that, at least for non script-like situations, there must be some salient entity in the preceding discourse to which the fragment stands in some kind of relation. 126 of the 216 cases turned to have an antecedent in the previous discourse. This amounts to the 58.3% of the total. In most fragments of this type the missing relation is an identity relation between the entity or property provided by the fragment and a salient entity in the context. However, sometimes, there is more than one salient entity in the context and it is knowledge of the world that tells us which is the one referred to by the fragment, like in the following example. Moreover, without the necessary knowledge one still does not know that what is meant is the title of the movie *A Room With a View*.

(2) MIG: Una habitación sin vistas?
A room without views
'(Is it) A room without views?'
CRI: No, es una habitación con vistas. Pero es yo qué sé...
No is a room with views. But is I what know
'No, it is a room with views, but it is, how would I say...'
PAT: En Cuatroca.
In Cuatroca
'(It is) In Cuatroca.'
MIG: Como la película.
Like the movie.
'(It is a room with views) Like (the title of) the movie.'

90 items (41.7%) did not have an antecedent in the preceding discourse. Within those we found some which do indeed refer to some salient entity in the context, not explicitly uttered in the discourse, but part of the surrounding environment. The corpus contains some situation and environment descriptions but they aren't part of the semantic tagging, so at the moment we cannot recover this kind of references. For example:

(3) **Situation:** Two friends looking at a shop window. One points to a necklace and says:

Qué chulo!
How cool
'(This necklace is) So cool!'

We also found cases where it is the situation which tells us how to resolve the fragment. In the following example the fragment is interpreted as 'Speak louder' but in a situation where somebody is hanging something on the wall, it should be interpreted as 'Hang it higher', since the Spanish word 'alto' means both 'high' and 'loud'.

(4) **Situation:** After having asked a question to the whole group, the teacher says:
Más alto.
More loud
'(Speak) Louder.'

Others, however, do not seem to have any relation with the context. These are cases of alone-standing gapping, where the omitted verbal predicate has never been uttered explicitly. It seems that the syntactic marking of the arguments allows them to be recognized as such and default basic predicates expressing identity, location or movement can be infered. However, as Spanish has no morphological case-marking the communicative context may sometimes play an important role when disambiguating between a location and a movement, for example.

(5) Junto a los números la interpretación.
Next to the numbers the interpretation
'Next to the numbers (is) the interpretation.'

For those fragments without an explicit linguistic source but which refer to some salient entity in the preceding discourse we calculated the distance in sentences between the fragment and the salient entity. The results are similar as for the fragments with an explicit linguistic source. In 67.08% of the cases the distance is 1 sentence; in 20%, 2 sentences; 4.5%, 3 sentences. Surprisingly, 2.5% of the cases are cataphoras, with a distance -1.

Finally, we also looked at the lemmas of the most frequently omitted relations. Not surprisingly, these are the lemmas 'ser', 'estar' (both translated as 'to be'), 'haber' (there is) and 'tener' (to have). The table shows the most frequent ones[4]:

[4]The annotation recovers the omitted lemma. Since this one hasn't been uttered explicitly, the most basic default lemma is chosen.

Lemma	total	percentage
ser	139	64.1
estar	48	21.7
haber	10	4.6
tener	4	1.8

3.3 Event parts

We have looked at the types of arguments filled by remnants and the results show a clear predominance of indirect relations. These are followed by patients and properties, that is, second arguments of events. Finally, entities, locations and agents are the less frequent argument types. This confirms our expectations that remnants tend to be of oblique argument types. The following table shows the results:

Event part	% Occurrences
Indirect relations	37.12%
Properties	26.65%
Patients	14.37%
Entities	11.98%
Locations	7.19%
Agents	2.69%

4 Discussion and Conclusion

In this study we have analysed two kinds of ellipsis - on the one hand, those having a linguistic source, and, on the other hand, those having no linguistic source. For the first group the most frequent distance from the source turned out to be 1 sentence. When looking at long distances, we have seen that the source is accessible because of its position in the discourse structure. However, when looking at more intermediate distances, like in the following example (3 sentences), we have sometimes found that the source is not in an accessible position in the discourse structure, but it is still available as the source of ellipsis.

(6) AS: Lleva un año fuera del sindicato CCOO. Por dónde ha orientado
Is a year out of the syndicate CCOO. For where have
ahora su vida? Porque decían que pretendía ser incluso presidente
orientated now your life? Because said that tried be even
del CES. Del Consejo Económico Social.
president of the CES. Of the Council Economic Social.
'You have been out of the CCOO syndicate for a year. Which direction have you given to your life? Because some said that you hoped to become even the president of the ESC. Of the Economic Social Council.'

REZ: Bueno fuera del sindicato no. Sigo como afiliado.
Well out of the syndicate not. Am as member.
'Well (I have) not (been) out of the syndicate. I am still there as a member'

A possible explanation for this phenomenon is that the syntactic structure of the source is still in working-memory and it can be accessed to resolve the ellipsis.

For ellipsis far away from the source and ellipsis without antecedent, we believe that there is no syntactic representation available in memory. In the first case there must be a semantic representation of the source because, while syntactic structure decays gradually, semantic structure may remain in memory depending on its degree of importance for the overall discourse or whether it is a topic not concluded yet.

Ellipsis without antecedent and refering to some salient entity in the context have some similarity with 0-anaphors. Their resolution involves to find a referent in working-memory, like in anaphora resolution and to infer some relation holding between the entity and the remnant. The cases of gapping without antecedent involve finding a relation holding between the several remnant constituents. It is logical to think that these are intrinsically semantic processes, since it is the meaning of the entities together with the meaning of the argument roles which helps to find the missing relation.

Even when the ellipsis has a source, it is sometimes not clear how the fragment should be resolved, that is, there are many surface form possibilities for the resolution, though it may be clear what it is meant or at least which are the ultimate goals which the fragment accomplishes, as illustrated in (1). This makes desirable to have a semantic representation (sometimes rather general) as the result of ellipsis resolution. (1) also shows that even when there is structural identity between source and fragment pragmatic reasoning may play an important role on resolving the ellipsis.

We would like to emphasize the importance of working with corpora when studying this kind of phenomena. Although there are very few resources (corpus annotated with coreference information), ellipsis is specially important in spontaneous spoken language and we think it is worth to study it.

References

Alcántara M.: 2005, *Anotación y recuperación de información semántica eventiva en corpus*, PhD thesis, Universidad Autónoma de Madrid.

Cooper, R. and J. Ginzburg: 2001, Resolving ellipsis in clarification, *Proceedings of the 39th Meeting of the Association for Computational Linguistics (ACL)*, Morgan Kaufman Publishers, San Francisco, pp. 236–243.

Cooper R., Engdahl E., Larsson S. and S. Ericsson: 2000, Accomodating questions and the nature of QUD., *Proceedings of the GÖTALOG*, pp. 57–61.

Grosz B. J., Joshi A. K. and S. Weinstein: 1995, Centering: A framework for modelling the local coherence of discourse, *Computational Linguistics* **21**(2), 203–225.

Hardt D. and M. Romero: 2004, Ellipsis and the structure of discourse, *Journal of Semantics* **21**(4).

Kintsch W. and T. A. van Dijk: 1978, Toward a model of text comprehension and production, *Psychology Review* **85**.

Moreno A., De la Madrid G., Alcántara M., González A., Guirao J. M. and R. D. la Torre: 2005, The Spanish corpus, *in* E. Cresti and M. Moneglia (eds), *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages*, John Benjamins.

Schlangen, D.: 2003, *A Coherence-Based Approach to the Interpretation of Non-Sentential Utterances in Dialogue*, PhD thesis, Institute for Communication and Collaborative Systems, School of Informatics, University of Edinburgh.

Streb J., Hennighausen E. and F. Rösler: 2004, Different Anaphoric Expressions are Investigated by Event-Related Brain Potentials, *Journal of Psycholinguistic Research* **33**(3).