Research on Language and Computation manuscript No.
(will be inserted by the editor)

Combining Multiple Information Sources for Ellipsis

Jennifer Spenader[1] **and Petra Hendriks**[2⋆]

[1] Artificial Intelligence,
University of Groningen, Groningen, The Netherlands
e-mail: J.Spenader@ai.rug.nl
[2] Center for Language and Cognition Groningen,
University of Groningen, Groningen, The Netherlands
e-mail: P.Hendriks@rug.nl

September 5, 2006

Abstract What knowledge sources are necessary in the interpretation and generation of ellipsis? After a short background on earlier approaches we compare and discuss each of the four papers selected for this special issue, examining how they approach ellipsis generation or interpretation. We highlight areas where more research needs to be done: outlining how pragmatics affects ellipsis, empirical studies, and theoretical work on what the effect of ellipsis is in context.

Key words ellipsis, ellipsis resolution, generation of ellipsis, cross-modular approaches, pragmatics

1 Introduction

Ellipsis has been a major topic in linguistics since the first formal analyses of natural language were developed. This fascination comes in part because its behavior and constraints on its use suggest the presence of hidden structures and necessitate theorizing about how this 'silence' is interpreted. There is still little consensus about how ellipsis should be analyzed, what its actual function is, or even the very basic question of what constructions belong to the category. But many exciting ideas are being currently debated and this is reflected in this issue.

⋆ Jennifer Spenader (grant no. 355.70.005) and Petra Hendriks (grant no. 015.001.103) gratefully acknowledge the Netherlands Organisation for Scientific Research, NWO, for financial support.

This special issue is a collection of revised papers originally presented at the 2005 European Summer School in Logic, Language and Information in Edinburgh, Scotland (Spenader and Hendriks 2005). That workshop, entitled "Cross-modular Approaches to Ellipsis", was intended to stimulate research into how different information sources, i.e. semantic, syntactic and pragmatic, contribute and, more importantly, interact in the interpretation and generation of elliptical utterances. Further, we strongly encouraged submissions that looked at empirical data, with an eye to encouraging research that would be useful when incorporating ellipsis into Natural Language Processing (NLP) applications. We specifically solicited contributions focusing on:

- implemented ellipsis resolution algorithms that incorporate information from more than one linguistic module
- appropriate generation of ellipsis
- studies of ellipsis in dialogue and the relation of ellipsis to discourse structure
- formalized treatments of ellipsis that incorporate semantic, pragmatic and discourse structural information
- corpus studies of elliptical phenomena
- elicitation tasks that give insights into interpretation or generation of elliptical phenomena

Our intentions were in many ways fulfilled, and this is reflected in the content of the papers that were chosen for this special issue. These papers all present original research that addresses several of the major issues being debated in ellipsis research today.

In this introduction we give an orientation to each of the four papers selected and explain how they address some of the major areas of controversy in current work on ellipsis. We also compare their approaches and results. Finally in the last section we discuss the research questions raised in our original call for papers that were not addressed at the workshop and which have not received much attention elsewhere. Here we see clear directions for future work that can help refine the ongoing debate.

2 Structure in Ellipsis?

One of the major questions currently debated in ellipsis research is whether or not ellipsis sites contain hidden syntactic structure. This fundamental question has divided researchers into two camps. Researchers such as Sag (1976), Hankamer (1979), Fiengo and May (1994) and Merchant (2001) believe there is such structure. In contrast, researchers such as Dalrymple et al. (1991), Hardt (1999), and more recently Dalrymple (2005) believe that it is possible to treat ellipsis as purely a semantic phenomenon.

But in the last few years several mixed proposals have been put forward. One of the first hybrid proposals was Kehler (2000), who argued that

syntax constrains ellipsis only when the construction containing ellipsis is related via a rhetorical relation of parallelism or contrast to the rest of the discourse. When the elliptical construction expresses another discourse relation such as explanation, syntactic constraints no longer limit the felicity of the ellipsis. In a recent paper, Kennedy (forth.) argues instead for a distinction based on the type of syntactic relation involved. Elided constituents, since they are not pronounced, are argued to be unaffected by syntactic constraints stemming from morphophonological properties. Other syntactic constraints, such as those involved in case theory, will however affect ellipsis and account for some of the restrictions on ellipsis. Several researchers have gone even further. Hendriks and de Hoop (2001) show that the interpretation of nominal anaphors and elliptical comparatives can be accounted for through the interaction of violable constraints that are syntactic, semantic as well as pragmatic in nature. These mixed proposals have moved away from a unary account that attempts to do everything with syntax, or everything with semantics, and instead are open for different linguistic modules influencing elliptical constructions at the same time. It was exactly this type of cross-modular interaction that was the focus of the ESSLLI workshop and hence of the contributions in this volume, some of which also incorporate other modules than the ones mentioned above.

This fundamental issue as to whether syntax or semantics alone is sufficent to explain elliptical behavior is directly addressed in Hoeksema (this issue). Hoeksema's paper, 'Pseudogapping. Its syntactic analysis and cumulative effects on its acceptability', reviews the previous syntactic approaches that all argued that pseudogapping was the result of movement followed by deletion. But Hoeksema thoroughly summarizes a number of contradictory empirical observations about pseudogapping and concludes that none of the current proposals can account for this body of facts. He instead suggests a pro-form semantic approach based on Miller (1990) and similar to work by Hardt (1999), which doesn't require movement and instead treats the realized *do* as an anaphor.

Hoeksema's paper comprehensively lists features that distinguish pseudogapping from gapping and VP-ellipsis that are unaccounted for in current research. But its main contributions are empirical ones. Hoeksema does both a corpus study of pseudogapping in English and a similar construction in Dutch, and a judgement study of native English speakers as to the felicity of pseudogapping sentences by presenting subjects with manipulated sentences with different relevant features, e.g. comparative or simply coordinating, or examples with or without the presence of a predicate remnant. Hoeksema found that pseudogapping prefers comparative contexts without remnants, and that examples with coordination and a remnant were judged much worse than examples that only contained one of these disprefered structures, which were in turn worse than pseudogapping examples with neither. Cumulative effects like these are not handled well in current approaches to syntax, but as Hoeksema points out have been found for gapping in work by Keller (2000) and Sorace and Keller (2005).

Theune et al.'s contribution 'Performing aggregation and ellipsis using discourse structures' technically takes a syntactic deletion approach to ellipsis, although the dependency trees these authors use should perhaps better be characterized as pre-syntactic representations since they do not express linear word order. Thus they follow the analysis first introduced by Sag (1976). Under this view, ellipsis is the deletion of lexical material by the speaker. However, the authors didn't choose the deletion approach because of deeply rooted theoretical conviction, but for more practical reasons – this method, not uncommon in NLG (Natural Language Generation) systems, works well for creating natural, less redundant text. Theune et al. do acknowledge that certain types of ellipsis are more sensitive to syntactic constraints such as island constraints than others, and therefore treat the various types of ellipsis as different constructions subject to different constraints.

Unfortunately many approaches to "aggregation", the NLG term for processes that, among other things, create elliptical utterances from non-elliptical ones, are often just mechanical steps within the generation process, and the hard questions of why and when ellipsis is grammatical, or even felicitious, are not answered. Our workshop desiderata was in part a reaction to this missing work. Theune et al.'s work fills this gap by going beyond merely mechanically manipulating syntactic structure. Their implementation uses rhetorical structure to constrain the felicity of the application of the deletion rules, an original contribution. In this way they incorporate discourse and contextual information. Their work includes an in depth study of rhetorical structure marking cue phrases in Dutch. In effect, syntactic aggregation in their system must be licensed by rhetorical structure. This is clearly a necessary constraint on syntactic aggregation, because, as results from work by Kehler (2000) and Hendriks (2004) have shown, eliding structures can sometimes void potential rhetorical interpretations that were possible in the unelided form. For example, applying gapping to two conjuncts that are ambiguously in a causal or parallel rhetorical relation removes the causal reading, which in many cases can be the intended reading.

However, rhetorical relations are still a very coarse tool for capturing discourse structure and not all types of reduced utterances can be generated by appealing to syntactic manipulations. Fragments or sub-sentential units cannot be generated this way. Ericsson's paper, 'Optimising elliptical utterances in dialogue', examines just these types of utterances, with a novel information structure-based method that captures contextual effects. Ericsson points out that naming these short utterances 'fragments' or 'sub-sentential units' gives the wrong connotations, as they are fully interpretable utterances within their context of use. She explicitly states that she follows Stainton (forth.) in not considering fragments to be fragmentary or derived by deletion. Her analysis incorporates the results of an empirical study of the use of short or elliptical utterances or fragments in naturally produced dialogue (Ericsson 2005). Having studied natural examples, Ericsson is then able to generate questions and answers in dialogue, making reference only to

information structural categories such as FOCUS, BASE and GROUND. This is an innovative pragmatic approach; few researchers have formalized in such detail the way in which the relation between the information content of an utterance and the discourse context affects the form of the utterance. The focus is placed on how speakers select what material can be left unarticulated while still preserving recoverability given the context. Ericsson does all of this in an optimality theory analysis, concentrating on non-syntactic constraints and persuasively demonstrating that felicitious ellipsis generation will need to make reference to information structure.

But are these short utterances a form of ellipsis? Their proper treatment is another major source of current debate, and recently a collection of papers was published on just this topic (Elugardo and Stainton 2005). Ericsson's examples share enough characteristics with traditional cases of ellipsis that it seems sensible to study them within this research. Further, since Ericsson's approach makes some specific predictions about when reduced forms are appropriate give a specific context, it seems a promising way to begin discussing the licensing conditions of more traditional forms of ellipsis.

Both Theune et al. and Ericsson look at ellipsis from the perspective of generation. Even though their objects of analysis and their intended results are very different, it is illuminating to compare them to each other. Theune et al. is work from the perspective of text generation, and looks at how syntactically described processes can be applied to make more natural sentences. The naturalness of the resulting sentences is believed to arise in part because the elided version removes redundancy. Could, however, this redundancy and its felicitous removal instead be described in the pragmatic terms of information structure of Ericsson's analysis? This might lead to an even more accurate account of why the aggregated sentences seem to be more natural. Further we know from corpus studies of ellipsis such as Meyer (1995) that ellipsis doesn't get applied in all cases where it structurally could. Some of these exceptions may be because of rhetorical constraints like the ones explored in Theune et al., but some may be for information structural reasons. How these pragmatic considerations interact with ellipsis are exciting questions for future research.

The fourth paper of this special issue, Repp's work '¬(A& B). Gapping, Negation and Speech Act Operators', also presents examples where pragmatic information systematically influences the interpretations available. Repp discusses the interaction of discourse constraints with semantic and syntactic ones, by arguing that whether the negation in gapped sentences takes wide scope over the entire coordination, or whether it is interpreted in each conjunct individually (= distributed scope) actually depends on the type of speech act that the gapped utterance is being used for. Wide scope negation is argued only to be available in denials or similar speech acts, and distributed scope in other cases. This is a particularly interesting analysis because it shows how pragmatic factors can explain an otherwise confusing set of syntactic and semantic facts. Repp's analysis, which looks closely at the phenomenon in German, further illustrates the importance of studying

ellipsis in a variety of languages, because the anomaly of the wide scope readings are most clear when other facts for German are considered.

3 'Elided' research?

There are several research questions still unaddressed or scantily addressed in the current body of research on ellipsis, yet many of them are obvious gaps to be filled.

All the papers in this special issue find problems with a purely syntactic approach, incorporating other information sources. Hoeksema rejects a purely syntactic approach to pseudogapping in favor of a semantic approach. Theune et al., Ericsson and Repp all incorporate some type of pragmatic information in their account of ellipsis. As for generation, Theune et al. and Ericsson offer insightful approaches to improving or accounting for reduced forms using rhetorical relations and information structure, respectively. Repp shows how the interpretation of negation in gapped sentences must be determined by the speech act expressed. Thus all papers present evidence of the necessity of incorporating non-syntactic knowledge, supporting cross-modular approaches to ellipsis. The relation of ellipsis to discourse structure is clearly there, and there is also a relationship to information structure. Unfortunately, currently these two lines of research are being performed by different research groups, with quite different goals.

Despite the large body of research done on ellipsis, very little of this research has an empirical basis. The corpus study of Meyer (1995), which looked at frequencies of conjunction reduction, gapping and right-node raising in English newspaper texts, is rather unique. In another study Alcantara and Bertomeu (2005) did a study of ellipsis in spontaneous speech in Spanish and found that 7.5 percent of sentences were elided. What most corpus studies have in common is that they are basically counting occurrences of ellipsis. Meyer (1995) also considers examples where ellipsis could have been used but was not, which sheds some light on speaker choices. But basically these are quantitative studies, and the actual function or contribution of the elliptical construction to the context isn't addressed in depth. Among our authors, Ericsson's paper does build on corpus work. Further, in addition to the corpus work, Hoeksema also looks at how real speakers (i.e. not linguists!) judge examples, giving us insights in the interpretation of elliptical expressions. Ideally the results of corpus work should always be part of an informed theory, but more such studies are needed to provide theorists with a stonger empirical base.

Empirical work is also necessary in order to incorporate ellipsis into NLP applications. Just like the misuse of anaphora has been shown in processing studies to lead to confusion and problems in comprehension, the misuse or non-use of ellipsis when it is appropriate might lead to processing difficulties for hearers or readers. But these types of processing studies haven't actually been done, even though several studies have investigated the processing of

ellipsis (e.g. Shapiro and Hestvik 1995, Frazier and Clifton Jr. 2000 and Frazier and Clifton Jr. 2001).

And we still have far to go in explaining what the function of ellipsis is. The choice of elliptical utterences over non-elliptical options is standardly argued to be motivated by speaker economy. However, Hendriks and Spenader (2005) list a number of cases where ellipsis has a different function. Ellipsis sometimes removes possible readings, indicates the default interpretation, or allows the speaker to express something that otherwise violates syntactic constraints. But how frequent are these functions? Research on ellipsis should look at function as much as it has already looked at form.

Despite all the theories and research that has been produced, we are still not much closer to determining what ellipsis actually does, and why in some cases a structurally possible elided form isn't used. Studies of ellipsis incorporating discourse structure and information structure, corpus work on ellipsis, and studies investigating the online processing of ellipsis may help reveal the actual function and contribution of ellipsis in natural language. This is surely necessary to improve applications for Natural Language Processing and Natural Language Generation.

4 Conclusion

In this paper we gave an orientation to the four papers selected for this special issue on ellipsis. All papers emphasize the need to incorporate non-syntactic information sources such as semantics, rhetorical relations, information structure and speech acts to account for the interpretation and generation of ellipsis. Incorporation of these information sources is in some of the papers shown to improve the use of ellipsis in NLG applications. We concluded with a list of open issues in the area of ellipsis, and gaps to be filled, which we hope will inspire others.

References

Alcantara, M. and N. Bertomeu: 2005, 'Ellipsis in Spontaneous Spoken Language'. In: J. Spenader and P. Hendriks (eds.): *Proc. of the ESSLLI Workshop on Cross-Modular Approaches to Ellipsis*.

Dalrymple, M.: 2005, 'Against Reconstruction in Ellipsis'. In: R. Elugardo and R. Stainton (eds.): *Ellipsis and Non-Sentential Speech*, Studies in Linguistics and Philosophy. Springer.

Dalrymple, M., S. M. Shieber, and F. C. N. Pereira: 1991, 'Ellipsis and higher-order unification'. *Linguistics and Philosophy* **14**(4), 399–452.

Elugardo, R. and R. Stainton: 2005, 'Introduction'. In: R. Elugardo and R. Stainton (eds.): *Ellipsis and Non-Sentential Speech*, Studies in Linguistics and Philosophy. Dordrecht: Springer, pp. 1–26.

Ericsson, S.: 2005, 'Information Enriched Constituents in Dialogue'. Ph.D. thesis, University of Gothenburg.

Ericsson, S.: this issue, 'Optimising Elliptical Utterances in Dialogue'. *Research on Language and Computation.*

Fiengo, R. and R. May: 1994, *Indices and Identity.* Cambridge Massachusettss: MIT Press.

Frazier, L. and C. Clifton Jr.: 2000, 'On bound variable interpretations:The LF-only hypothesis'. *Journal of Pyscholinguistics Research* **29**, 125–139.

Frazier, L. and C. Clifton Jr.: 2001, 'Parsing Coordinates and Ellipsis: Copy α'. *Syntax* **4**(1), 1–22.

Hankamer, J.: 1979, *Deletion in coordinate structure.* New York: Garland Publishing, Inc.

Hardt, D.: 1999, *VPE as proform: some consequences for binding*, Empirical Issues in Formal Syntax and Semantics 2: Selected Papers from the Colloque de Syntaxe et Semantique a Paris. The Hague: Thesus.

Hendriks, P.: 2004, 'Coherence Relations, Ellipsis and Contrastive Topics'. *Journal of Semantics* **21**(12), 132–154.

Hendriks, P. and H. de Hoop: 2001, 'Optimality Theory Semantics'. *Linguistics and Philosophy* **24**, 1–32.

Hendriks, P. and J. Spenader: 2005, 'Why be silent? Some functions of ellipsis in natural language'. In: J. Spenader and P. Hendriks (eds.): *Proc. of the ESSLLI Workshop on Cross-Modular Approaches to Ellipsis.*

Hoeksema, J.: this issue, 'Pseudogapping. Its syntactic analysis and cumulative effects on its acceptability'. *Research on Language and Computation.*

Kehler, A.: 2000, 'Coherence and the resolution of ellipsis'. *Linguistic and Philosophy* **23**, 533–575.

Keller, F.: 2000, 'Gradience in Grammar:Experimental and Computational Aspects of Degrees of Grammaticality'. Ph.D. thesis, University of Edinburgh, Edinburgh, Scotland.

Kennedy, C.: forth., 'Ellipsis and Syntactic Representation'. In: K. Schwabe and S. Winkler (eds.): *The Syntax-Semantics Interface: Interpreting (Omitted) Structure.* John Benjamins.

Merchant, J.: 2001, *The Syntax of Silence: Sluicing, Islands and the Theory of Ellipsis.* Oxford: Oxford University Press.

Meyer, C. F.: 1995, 'Coordination Ellipsis in Spoken and Written American English'. *Language Sciences* **17**.

Miller, P.: 1990, 'Pseudogapping and *do so* substitution'. In: *Proceedings of CLS 26.* Chicago, pp. 293–305.

Repp, S.: this issue, '¬(A & B). Gapping, Negation and Speech Act Operators'. *Research on Language and Computation.*

Sag, I.: 1976, 'Deletion and logical form'. Ph.D. thesis, Massachusetts Institute of Technology.

Shapiro, L. P. and A. Hestvik: 1995, 'On-line comprehension of VP-ellipsis: Syntactic reconstruction and semantic influence'. *Journal of Psycholinguistic Research* **24**, 517–532.

Sorace, A. and F. Keller: 2005, 'Gradience in Linguistic Data'. *Lingua* **115**(11), 1497–1524.

Spenader, J. and P. Hendriks (eds.): 2005, 'Cross-Modular Approaches to Ellipsis'. Edinburgh, Scotland: ESSLLI Workshop.

Stainton, R. J.: forth., 'Neither fragments nor ellipsis'. In: L. P. et al. (ed.): *The Syntax of Nonsententials*. Philadelphia: John Benjamins, pp. 93–116.

Theune, M., F. Hielkema, and P. Hendriks: this issue, 'Performing Agggregation and Ellipsis Using Discourse Structure'. *Research on Language and Computation*.