Chapter 6

Verb phrase ellipsis and sloppy identity: a corpus-based investigation

Johan Bos
University of Groningen

In a corpus-based confrontation between sloppy and strict identity in elliptical contexts, the former beats the latter with a striking 9–0. Whether this results is representative for verb phrase ellipsis in general is a question of debate. Perhaps the sloppy players had home advantage, and strict players perform better in corpora other than the Wall Street Journal.

1 Introduction

In this article[1] I will present corpus instances of Verb Phrases Ellipsis (VPE), a linguistic phenomenon that manifests itself in the English language when an auxiliary verb is used to refer to a complete verb phrase mentioned elsewhere in the linguistic context, as in *Bill wrote a paper, and John did too.* Recently, Jennifer Spenader and myself annotated a large corpus of English newspaper text (parts of the Wall Street Journal) on occurrences of VPE (Bos & Spenader 2011). We undertook this effort because, until then, detailed annotation work of VPE carried out on a large scale did not exist, with the exception of Hardt (1997) and Nielsen (2005). The primary aim of this enterprise was to develop benchmark tools for automated ellipsis recognition and resolution in the context of natural language processing. However, the results can also be used to study the distribution and frequency of the various types of VPE and problems they trigger known from the rich linguistic literature on ellipsis.

Some of the findings of Bos & Spenader (2011) were expected, and some unexpected. Not surprising was the relative rarity of the phenomenon of VPE in newswire: on average they found only one instance of VPE in every 109 sentences (Bos & Spenader 2011). However, much to my surprise, is the lack of overlap of types of VPE found in

[1] This paper is dedicated to John Nerbonne, who introduced me more than twenty years ago to the world of computational semantics (Nerbonne 1996). He supervised my master's thesis work with an enormous amount of enthusiasm and expertise. This resulted in my first international publication (Bos 1993). I am proud to say that I was John's very first graduate student in Groningen.

the Wall Street Journal corpus with the classical examples found in the theoretical ellipsis literature. Well-studied phenomena such as pseudo-gapping, split antecedents, cascaded ellipsis, antecedent-contained deletion are scarce in the newspaper genre. So are VPE that give rise to what Dahl (1973) called *sloppy identity*.[2] It is this latter phenomenon that I will closer inspect in this article. Its statistical presence in real corpora was unknown, and it still is. In this article I will look at the occurrence of sloppy identity in the well-known Wall Street Journal corpus.

2 Sloppy identity

Many linguists are fascinated by the notorious strict/sloppy ambiguity that manifests itself in VPE (Dahl 1973; Sag 1976; Klein 1987). It occurs when the subjects of the source and target clause denote different entities, and a pronoun appears in the source clause and co-refers with the subject of the same clause. An example is *John likes his mother, and Bill does too*, where the strict interpretation gives rise to the reading where Bill likes John's mother and the sloppy interpretation yields the reading where Bill likes his own mother. Another example is *John has never read a Russian novel he disliked. But Bill has. It was War and Peace*,[3] taken from Gawron, Nerbonne & Peters (1991), where the strict interpretation is implausible because that would imply that John disliked a Russian novel that he never read.

This is without any doubt an interesting kind of ambiguity, and many computational solutions have been proposed for it (Dalrymple, Shieber & Pereira 1991; Bos 1994; Crouch 1995; Bos 2012). The question is how important this phenomenon is from a language technology point of view. To answer this question, I think it is good to look at naturally occurring data rather than examples invented by theoretical linguists. From the 554 cases of VPE that Bos & Spenader (2011) annotated in their one-million-word corpus, only *nine* show a *potential* ambiguity between a strict and sloppy interpretation. That is very little, perhaps even disappointingly little given the amount of theoretical work on the topic. What I am going to do in this article is to carefully study the behaviour of these nine cases. The main goal is to see whether the so-called sloppy identity is the rule or rather the exception.

3 Sloppy vs. strict in the Wall Street Journal corpus

Before I present the instances of VPE found by Bos & Spenader (2011), let me first introduce some notational conventions that I will use. For each occurrence of VPE, the antecedent VP is marked by square brackets, and the auxiliary verb triggering the elliptical VP is set in bold face. The pronoun causing the potential strict/sloppy ambiguity is underlined. Co-referential phrases are indicated by printing the same indices (i, j) in subscript. The cases are listed in the order they appear in the Wall

[2] Dahl attributes the origin and name of the problem to J. R. Ross.

[3] Incidentally, this example is an instance of the *Missing Antecedent Problem*, because the antecedent of the neuter pronoun is not explicitly available in the discourse.

Street Journal corpus. All of them are presented with a reference to the location in the corpus, which is the name of the raw file as distributed by the Penn Treebank (Marcus, Santorini & Marcinkiewicz 1993).

3.1 Carlos Saul Menem (sloppy vs. strict: 1–0)

The first instance we look at shows an odd kind of elliptical construction for a couple of reasons.[4] First, we have a present participle form of *do*, which is rather unusual for elided VPs. Second, there is a semantic mismatch between the parallel elements of the source and target clause: The subject of the source clause is a country (Brazil), whereas the subject of the target clause is a person (the President of Argentina). Here it is:

> **Case 1: Carlos Saul Menem** `<wsj_0415>`
> If Brazil$_i$ devises an economic strategy allowing it to resume growth and service debt, this could lead it$_i$ to [$_{\text{VP}}$ open up and deregulate its$_{i,j}$ sheltered economy], analysts say, just as Argentinian President Carlos Saul Menem$_j$ has been **doing** even though he was elected on a populist platform.

The presence of the possessive pronoun *its* in the source clause, referring to Brazil, causes the potential strict/sloppy ambiguity. However, the target clause certainly doesn't mean that the Argentinian president opened up Brazil's economy — that would be highly unlikely — hence a strict interpretation is out of the question. The sloppy interpretation of *its* yields Argentina as antecedent, of course. Note however, that this antecedent isn't overtly expressed in the text, which is a further reason why this case is interesting.

3.2 IBM (sloppy vs. strict: 2–0)

This is very much like a standard textbook occurrence of VPE, where the source and target are connected via a temporal adverb:

> **Case 2: IBM** `<wsj_0445>`
> IBM$_i$, though long a leader in the Japanese mainframe business, didn't [$_{\text{VP}}$ introduce its$_{i,j}$ first PC in Japan] until five years after NEC$_j$ **did**, and that wasn't compatible even with the U.S. IBM standard.

The source clause contains the possessive pronoun *its* that co-refers with the subject, *IBM*. Hence, there is a potential ambiguity in the target clause. The strict variant

[4] Daniel Hardt, in email correspondence on 17-04-2007, says the following about this example: "This is a variant of VPE that has not been much studied as far as I know. I think that the *as* binds a variable standing for a VP meaning, very much like a wh-operator, as you could have for example in a variant of the above *...which$_i$ Argentinian President Carlos Saul Menem has been **doing**$_i$ even though he was elected on a populist platform.* I think this would suggest that the missing VP should be linked up to the VP which *as* is modifying, in this case *open up and deregulate its sheltered economy.* And that is the reading I get."

would be paraphrased as *until five years after NEC introduced IBM's PC in Japan*, and the sloppy one as *until five years after NEC introduced NEC's PC in Japan.* Of course, only the latter, sloppy interpretation makes sense — after all, why would one introduce a product of one's competitor?

3.3 Mr. Engelken (sloppy vs. strict: 3–0)

Here we have an instance of a *do-the-same* type of VP anaphor. Interesting here is that the subject of the source clause denotes a plural entity, whereas the subject of the target clause is a singular noun phrase:

> **Case 3: Mr. Engelken** `<wsj_0758>`
> Some 34,320 fans$_i$ jammed the stands, and [$_{\text{VP}}$ shouted at the top of $\underline{\text{their}}_{i,j}$ lungs]. Mr. Engelken$_j$ was **doing the same** across the Hudson River in New Jersey, where, with his nose pressed against the front window of the Passaic-Clifton National Bank, he watched the duel on a television set the bank set up for the event.

The potential strict/sloppy ambiguity here is caused by the plural possessive pronoun *their*. It is physically impossible to shout at the top of someone else's respiratory organs, at least in the preferred, non-literal sense of 'at the top of someone's lungs' that is obviously used here. Hence, there is no ambiguity here at all and the pronoun needs to be sloppily interpreted to get the desired reading. Note that the mismatch in number of the parallel subjects doesn't seem to matter at all to get the interpretation that Mr. Engelken was shouting at the top of his lungs.

3.4 Mr. Lawson (sloppy vs. strict: 4–0)

In order to fully comprehend the following case it might help to provide some context. It's 1989, we're in the UK. John Major has just been appointed Chancellor of the Exchequer, succeeding Nigel Lawson. Margaret Thatcher was Prime Minister at the time, as well as leader of the Conservative Party. Now consider:

> **Case 4: Mr. Lawson** `<wsj_0883>`
> Neil Kinnock, Labor Party leader, dubbed the 46-year-old Mr. Major$_i$ a "lap dog" unlikely to [$_{\text{VP}}$veer from $\underline{\text{his}}_{i,j}$ boss's strongly held views], as Mr. Lawson$_j$ sometimes **did**.

Again we have a possessive pronoun, *his*, in the source clause, that co-refers with *Mr. Major*.[5] So Mr. Major isn't likely to veer from the views of his boss, Margaret Thatcher. And Mr. Lawson? He sometimes *did* veer from his boss's views, but of

[5] We note in passing that this is, in fact, an interesting control construction, as the subject of the source clause isn't explicitly expressed.

course the only sensible way this makes sense is that this be his own boss, not Mr. Major's. Now, it turns out the case that Mr. Major and Mr. Lawson had in fact the same boss, namely Margaret Thatcher. Hence, extensionally speaking, the strict and sloppy reading would both lead to the same interpretation anyway.

3.5 Mr. Turner (sloppy vs. strict: 5–0)

This case features the broadcasting company Comsat Video wishing to contract the Denver Nuggets Basketball team. Comsat Video happens to be a rival of Turner Broadcasting System Inc. Here is the example:

> **Case 5: Mr. Turner** `<wsj_1461>`
> Comsat Video$_i$, which distributes pay-per-view programs to hotel rooms, plans to [$_{\text{VP}}$ add Nuggets games to their$_{i,j}$ offerings], as Mr. Turner$_j$ **did** successfully with his Atlanta Hawks and Braves sports teams.

Once more we have a possessive pronoun causing a potential ambiguity. Obviously, Mr. Turner didn't add the Atlanta Hawks and Braves to the offerings of Comsat Video, but to his own company. Hence, only the sloppy interpretation makes sense here.

We are now halfway in discussing the potentiallly ambiguous VPE cases. So far they have been all sloppy — is there a chance for a strict reading? Let's see…

3.6 Americans (sloppy vs. strict: 6–0)

This is an example similar in structure and analysis to a case we considered earlier: a temporal adverb connecting the source with the target clause, and a possessive pronoun causing a potential strict/sloppy ambiguity:

> **Case 6: Americans** `<wsj_1591>`
> "What this means is that Europeans$_i$ will [$_{vp}$ have these machines in their$_{i,j}$ offices] before Americans$_j$ **do**," the spokesman said.

As it is absurd to think that Americans have machines in offices of Europeans, only the sloppy interpretation of the possessive pronoun *their* is available.

3.7 Democrats (sloppy vs. strict: 7–0)

Here we have a comparative construction coinciding with VPE. Both GOP senators and Democrats turn back a percentage of their allocated personal staff budgets.

> **Case 7: Democrats** `<wsj_1695>`
> First, economists James Bennett and Thomas DiLorenzo find that GOP senators$_i$ [$_{vp}$ turn back roughly 10% more of their$_{i,j}$ allocated personal staff budgets] than Democrats$_j$ **do**.

The strict interpretation would yield an interpretation where Democrats turn back allocated personal staff budgets of GOP senators, which isn't a realistic possibility. Hence the sloppy interpretation of the possessive pronoun is the only way to interpret it.

3.8 Most magazines (sloppy vs. strict: 8–0)

This is an interesting instance of VPE, because the subject of the source clause isn't overtly expressed. The target clause is recovered *as most magazines spread out ads among its articles*, where *its* co-refers, obviously, with the subject of the target clause, *most magazines*, not to National Geographic. In other words, once more we end up with a sloppy interpretation:

> **Case 8: Most magazines** `<wsj_2109>`
> Another sticking point for advertisers was National Geographic$_i$'s tradition of lumping its ads together, usually at the beginning or end of the magazine, rather than [$_{\text{VP}}$ spreading ads out among its$_{i,j}$ articles], as most magazines$_j$ **do**.

This case comprises, in addition, a complex nominalisation controlling the subject of the coordinated present participle constructions *lumping its ads together* and *spreading ads out among its articles.*

3.9 Competitors (sloppy vs. strict: 9–0)

This last case involves another comparative form of VPE, escorted by subject-auxiliary inversion in the target clause. The potential ambiguity is caused by a third-person plural pronoun. It is, in fact, an example of a "lazy" pronoun (Geach 1962), because it stands for a literal repetition of a full definite noun phrase:

> **Case 9: Competitors** `<wsj_2109>`
> But the magazine was slower than its competitors to come up with its regional editions, and until last year [$_{\text{VP}}$ offered fewer of them] than **did** competitors.

The subject of the source clause is *the magazine*, which refers to National Geographic. The pronoun *them* refers to regional editions of the National Geographic. Hence, the items of comparison are the number of regional editions of the National Geographic offered by the National Geographic, and the number of regional editions of competitors (not the number of regional editions of the National Geographic, that would be ridiculous). Hence, apart from the complexity introduced by the comparative and the lazy pronoun, the analysis shows that, once again, we have a sloppy interpretation on our hands.

4 Discussion

The much discussed ambiguity between strict and sloppy interpretation caused by VPE is actually very rare in newswire text. Of more than 500 cases of VPE found in

the Wall Street Journal corpus, only nine cases showed potential sloppy/strict ambiguity. It turns out that all nine of them unequivocally show sloppy identity. It is true that we are working with a relatively small dataset of a restricted domain. Yet it is striking.

In eight out of nine cases a possessive pronoun caused the potential ambiguity. The balance between singular and plural number was equal: four in each case. The remaining case was a lazy pronoun referring to a possessive noun phrase. In the literature on VPE often examples with personal pronouns are given, but we found none in our corpus study.

Interestingly, several cases of surface semantic agreement conflicts were encountered. The Carlos Saul Menem example shows disagreement in the parallel subjects of the source and target clause. The Mr. Engelken example shows a mismatch in number between the source and target interpretation of the elided VP. Three of the VPE instances involve (complex) control constructions.

5 Conclusion

One could argue that any conclusion drawn from this dataset isn't significant because of its relatively small size. After all, we're only talking about *nine* examples. For the sake of the argument, let's assume that the distribution of strict and sloppy interpretations would be equally divided in texts. The odds to draw nine instances of VPE with a potential strict/sloppy ambiguity from a large corpus, which then turn out to be all of sloppy identity, are really low. So it is very likely that there is no equal distribution between sloppy and strict interpretation — informal Google searches confirm this claim.

It is certainly true that we need more empirical work and we should inspect larger and other genres of text. I have only looked at a very specific text genre: newswire. Journalistic prose is often associated with competent writing and governed by style guides. Often, texts are heavily edited for the sake of clarity and readability. In the case of the Wall Street Journal, its style guide gives the general advice to avoid ambiguity, however without saying anything in particular on pronouns and ellipsis (Martin 2002). Undeniably, empirical work on ellipsis should be extended to cover other genres of text, including spoken dialogue.

My tentative conclusion is that we don't need sophisticated algorithms in language technology, whose practitioners are content with accuracy figures of 90% or more given the inherent difficulty of the task, to compute all strict and sloppy interpretation for instances of VPE. First of all, because it is an extremely rare phenomenon, and secondly because if one defaults on sloppy identity a high accuracy is achieved already. As a consequence, computational implementations of ellipsis resolution algorithms could be far simpler than assumed so far. However, they could be more complicated with respect to other linguistic aspects, such as coordination, control, and mismatch between parallel elements.

References

Bos, Johan. 1993. VP ellipsis in a DRT-implementation. In *Proceedings of the sixth Conference of the European Chapter of the ACL (student session)*, 425–430. Utrecht, Netherlands.

Bos, Johan. 1994. Presupposition & VP ellipsis. In *Proceedings of the 15th International Conference on Computational Linguistics (Coling 1994)*, 1184–1190. Kyoto, Japan.

Bos, Johan. 2012. Robust VP ellipsis resolution in DR Theory. In Staffan Larsson & Lars Borin (eds.), *From quantification to conversation*, vol. 19 (Tributes), 145–159. College Publications.

Bos, Johan & Jennifer Spenader. 2011. An annotated corpus for the analysis of VP ellipsis. *Language Resources and Evaluation* 45(4). 463–494.

Crouch, Richard. 1995. Ellipsis and quantification: a substitutional approach. In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, 229–236. Dublin, Ireland.

Dahl, Östen. 1973. On so-called sloppy identity. *Synthese* 26. 81–112.

Dalrymple, Mary, Stuart M. Shieber & Fernando C. N. Pereira. 1991. Ellipsis and higher-order unification. *Linguistics and Philosophy* 14. 399–452.

Gawron, Jean Mark, John Nerbonne & Stanley Peters. 1991. The absortion principle and E-type anaphora. In. *Proceedings of the 2nd Conference on Situation Theory and its Applications*. J. Mark Gawron, Gordon Plotkin & Syun Tutiya (eds.). Stanford. 335–362.

Geach, P. T. 1962. *Reference and generality: an examination of some medieval and modern theories*. Cornell University Press.

Hardt, Daniel. 1997. An empirical approach to VP ellipsis. *Computational Linguistics* 23(4). 525–541.

Klein, Ewan. 1987. VP ellipsis in DR Theory. *Studies in Discourse Representation Theory and the Theory of Generalised Quantifiers*. Jeroen Groenendijk et al. (eds.).

Marcus, M. P., B. Santorini & M. A. Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics* 19(2). 313–330.

Martin, Paul R. 2002. *The Wall Street Journal essential guide to business style and usage*. Wall Street Journal Books.

Nerbonne, John. 1996. Computational semantics–linguistics and processing. In Shalom Lappin (ed.), *Handbook of contemporary semantic theory*, chap. 17, 461–484. London: Blackwell Publishers.

Nielsen, Leif Arda. 2005. *A corpus-based study of verb phrase ellipsis identification and resolution*. King's College London PhD thesis.

Sag, Ivan. 1976. *Deletion and logical form*. MIT PhD thesis.