The Distribution of Weak and Strong Object Reflexives in Dutch

Gosse Bouma and Jennifer Spenader

Information Science
University of Groningen
g.bouma@rug.nl

Artificial Intelligence
University of Groningen
j.k.spenader@rug.nl

Abstract

We use a syntactically annotated corpus to study the distribution of strong and weak reflexive objects in Dutch. Whereas previous work was limited to a small set of accidental reflexive verbs, we look at all transitive verbs in the corpus. We use subcategorization frames to approximate verb senses. We show that comparing the rate of pronominal usage to reflexive usage is a better predictor of strong or weak reflexive choice tendencies (giving a correlation of 33%) than considering all objects, confirming a suggestion by Haspelmath (2004). We also show that the automatic method gives results comparable to those for the semi-automatically collected data in Hendriks, Spenader, and Smits (2008).

1 Introduction

If a verb is used reflexively in Dutch, two forms of the reflexive pronoun are available. This is illustrated for the third person form in the examples below.

(1) a. Brouwers schaamt **zich/*zichzelf** voor zijn schrijverschap.
Brouwers is ashamed of his writing
b. Duitsland volgt **zichzelf** niet op als Europees kampioen.
Germany does not succeed itself as European champion
c. Wie **zich/zichzelf** niet juist introduceert, valt af.
Everyone who does not introduce himself properly, is out.

The choice between *zich* and *zichzelf* depends on the verb. Generally three groups of verbs are distinguished. Inherent reflexives are claimed to never occur with a non-reflexive argument, and as a reflexive argument are claimed to use *zich*

exclusively, (1a). Non-reflexive verbs seldom, if ever occur with a reflexive argument. If they do however, they can only take *zichzelf* as a reflexive argument (1b). Accidental reflexives can be used with both *zich* and *zichzelf*, (1c). Accidental reflexive verbs vary widely as to the frequency with which they occur with both arguments and it is this distribution that we would like to explain.

What exactly governs the choice between the weak and strong forms of a reflexive in the case of accidental reflexive verbs is largely unclear. The influential theory of Reinhart and Reuland (1993) explains the distribution as the surface realization of two different ways of reflexive coding. An accidental reflexive that can be realized with both *zich* and *zichzelf* is actually ambiguous between an inherent reflexive and an accidental reflexive (which always is realized with *zichzelf*). An alternative approach is that of Haspelmath (2004), Smits, Hendriks, and Spenader (2007), and Hendriks, Spenader, and Smits (2008), who have claimed that the distribution of weak vs. strong reflexive object pronouns correlates with the proportion of events described by the verb that are self-directed vs. other-directed.

In this paper we investigate to what extent a broad corpus investigation provides evidence for this claim. For each verb sense, we count how often it occurs with a strong or weak reflexive, or with another object. As many verbs occur rarely with a reflexive, a large amount of (parsed) data is required. We use a 470 M word Dutch corpus, syntactically analyzed using the Alpino-parser (van Noord, 2006) and use the results to make observations about reflexive use in general, the utility of large, parsed data sets, as well as the limits of a purely syntactic, unsupervised approach.

2 Previous Work

Haspelmath (2004), Smits, Hendriks, and Spenader (2007), and Hendriks, Spenader, and Smits (2008) have claimed that the distribution of weak vs. strong reflexive object pronouns (i.e. reflexives that are the object of a verb) correlates with the proportion of events described by the verb that are self-directed vs. other-directed. The claim is that if a verb is rarely used to express self-directed events, there will be a tendency to use the strong reflexive form when it is used reflexively to signal this marked use of the verb. The assumption behind the claim is that when the expectation that a given action will be self-directed is weak, emphasis on the reflexive argument is preferred, so the strong reflexive is used. Such emphasis is less likely if the verb is used with a self-directed meaning relatively often, and therefore the weak reflexive, which is shorter and should otherwise always be preferable, will be sufficient. This is in line with the claim that inherent reflexives

only occur with weak reflexives, since they only occur with reflexive meaning.[1]

Our research builds upon the work in Smits, Hendriks, and Spenader (2007) and Hendriks, Spenader, and Smits (2008), who studied the distribution of reflexive vs. nonreflexive use and the choice for a weak or strong form for 45 Dutch transitive verbs. Smits, Hendriks, and Spenader (2007) found a linear correlation between reflexive and non-reflexive usage (counting all third person NPs) for 21 % of the data in an 80 M word corpus (parsed using Alpino) for the verbs sufficiently frequent in the corpus. By combining this with judgement data, they were able to obtain an 83% correlation. Hendriks, Spenader, and Smits (2008), using a 300 M word corpus and 32 verbs obtained a correlation of 28% and a correlation of 30% when first and second person reflexives were included. Haspelmath (2004) suggests that only the ratio of pronominal objects to reflexive objects is relevant for determining the degree to which a verb is introverted (tends to describe self-directed events) or extroverted (tends to describe other-directed events). Hendriks, Spenader, and Smits (2008) found that the model proposed by Haspelmath yielded a correlation of 45%. However, they had no explanation as to why counting pronominal objects only gave more accurate results.

The research reported below differs from the approach of Hendriks, Spenader, and Smits (2008) in that we attempt to first empirically identify accidental reflexive verbs among all verbs in the corpus, and then use this very large set to test the different models of reflexive choice. The larger set of verbs may give us a more complete picture, but also forces us to adopt a fully automatic method for data collection, as we cannot afford to judge data individually for errors or unintended readings. In general, different senses of a verb may have very different tendencies for being used with self-directed activities. We therefore distinguish verbs by their different subcategorization frames in order to approximate verb senses.

3 Data Collection

We are interested in frequency estimates of the reflexive vs. nonreflexive use of the set of accidental reflexive verbs. Distinguishing accidental reflexives from inherent reflexives and non-reflexives is therefore crucial. A major problem is that most verbs are extremely ambiguous and simply checking if a verb can be used with a nonreflexive object or not is not sufficient:

[1] Note however that many inherent reflexives, like *zich herinneren*, (to remember) or *zich verspreiden*, (to spread out), can't really be characterized as being self-directed actions because the reflexive object doesn't seem to have a thematic role.

(2) a. De bedrijven maakten foute rekeningen op
The companies produced wrong bills
b. De schelpdieren maken al het voedsel op
The shellfish take all the food
c. Als ik 240 rijd, kan mijn assistente zich rustig opmaken
If I drive 240, my assistent can still put make-up on
d. De showbizz maakt zich op voor het huwelijk van het jaar
The showbizz prepares itself for the marriage of the year

The senses of *opmaken* illustrated in (2a) and in (2b) can hardly be used reflexively, the sense in (2c) can easily be used with a reflexive, while the sense in (2d) is inherently reflexive. Obviously, counting the frequency with which a verb occurs with an nonreflexive or reflexive object, without taking these differences in meaning into account, leads to noisy results. On the other hand, the parser does not annotate word senses, so we cannot automatically produce counts per verb sense.

The lexicalist nature of the Alpino-grammar implies that detailed verbal subcategorization frames are used to determine which complements a verb can combine with. By taking subcategorization frames into account some word sense distinctions can be identified. The inherent reflexive use of *opmaken* (2d), for instance, can be distinguished from the other senses by the fact that it subcategorizes for a PP-complement headed by the preposition *voor*.

Collecting counts for each pair of a verbal root + subcategorization frame is more precise than collecting counts per verbal root, but is still imperfect, as it fails to distinguish between verbal word senses with identical subcategorization frames. Verbs that have both an inherent reflexive use and an accidental reflexive use, for instance, are still problematic. (3a) illustrates a, highly frequent, idiomatic use of the verb *bedruipen*, which is inherently reflexive. Its meaning is clearly different from, although perhaps related to, the normal transitive use of *bedruipen* in (3b) (which is hardly found in the corpus).

(3) a. De vereningen kunnen zich met sponsoring bedruipen
The organisations can support themselves with sponsorships
b. Hij bedruipt een geitenkaasje met tijmhoning
He drips honey on a goat cheese

If *bedruipen* occurs with a reflexive, the parser has to choose between two verbal subcategorization frames: inherent reflexive or ordinary transitive. This choice is difficult, especially if the verb occurs with *zichzelf*. The inherent reflexive use is far more frequent than the ordinary transitive use. Nevertheless, in the case of *zichzelf*, the parser has a preference for using the ordinary transitive subcategoriza-

tion frame, instead of the frame associated with the inherent reflexive use.[2] This is unsurprising: strong reflexives in general do not occur with inherent reflexives. However, in ambiguous cases like this, this preference leads to inaccurate data. To avoid this problem, we discarded counts for all verb+subcategorization frames for which the parser has an alternative that differs from the current pair only w.r.t. the question whether the object obligatorily has to be a reflexive or not. This means that approximately 20% of the data is discarded.

Finally, we also decided to skip all occurrences of verbs that are used in passive sentences, or as complement of *laten*.

(4) a. De opstandelingen werden ontwapend
The rebels were disarmed
b. De kinderen laten zich niet dwingen
The children do not let themselves be forced

In passives, the object of the main verb appears as the subject of the passive auxiliary. In this position reflexives cannot be used. In sentences with *laten*, a reflexive may appear as the object of the embedded verb. This reflexive is interpreted as coreferential with the subject of *laten*, but it is unclear if it is also coreferential with the (unexpressed) subject of the embedded verb.

We used the 470 M word Twente News Corpus (TwNC), made up of the text of Dutch newspapers from the period 1994-2005 (Ordelman et al., 2007), which was parsed automatically with the Alpino-parser. Using the technology described in Bouma and Kloosterman (2007), we searched the corpus exhaustively for all occurrences of a verb with an object and a third person subject, and registered whether the object was *zich, zichzelf*, a (non-reflexive) pronoun, or a regular NP. We extracted 12 M verb-object tuples.

4 Distribution of Zich and Zichzelf

For accidental reflexive verbs in general, the use of *zich* was more frequent than *zichzelf*. We find 163K (84%) occurrences of *zich* vs. 31K (16%) occurrences of *zichzelf*. For more detailed observations, we restrict attention to verb+subcategorization pairs, that occur at least 50 times in the corpus, and at least 10 times with a reflexive (899 cases, of which, according to the grammar, 163 are inherent reflexive verbs, and 736 are accidental reflexive verbs). Although *zichzelf* in general is rare, we find that 6% of the accidental reflexive verbs (44 of 736) , when used reflexively, occur with a strong reflexive more than 95% of the time. Examples are *zichzelf in*

[2]Manual inspection of a sample suggests that in all uses of *zichzelf bedruipen* involve the *support oneself* meaning.

de weg zitten (hinder oneself), toespreken (address), opvoeren als (present), afschrijven (write off), and *onderbreken (interrupt)*. 34% of the accidental reflexive verbs (247) occur with a strong reflexive more than 50% of the time. 25% of the accidental reflexive verbs (187) occur with a strong reflexive less than 8% of the time. Some examples of the latter group are *beheersen (withhold), voorstellen (introduce), manoeuvreren (manoevre), uitleveren (hand over to), bevrijden (liberate), wassen (wash), (dress), scheren (shave), beschikbaar stellen (make available)*. We do find a number of 'outward directed' verbs among the group of verbs with a strong preference for *zichzelf*, and a number of 'self directed' verbs in the group with a dispreference for *zichzelf*. This is in line with Haspelmath's semantic characterization of such verbs.

The 44 verbs with a strong preference for the strong reflexive *zichzelf* were used non-reflexively 97.1% of the time. The 247 verbs used more often with a strong reflexive than with a weak reflexive were used non-reflexively 95.1% of the time. The 187 verbs used with a strong reflexive less than 8% of the time were used non-reflexively 72.0% of the time. This suggests that there is indeed a relationship between preference for the strong reflexive form and a high relative frequency of non-reflexive use.

Traditionally, it is claimed that inherent reflexives never occur with the strong reflexive *zichzelf*. We can examine empirically whether or not this is in fact true. Of the 163 reflexive verbs in our data-set, 112 (68.7%) occur with *zich* more than 99% of the time (often with only 1 or 2 occurrences of *zichzelf*).

The remaining 51 reflexive verbs occurred with strong reflexive objects more frequently. Here are a number of examples:

(5) a. Nederland moet stoppen zichzelf op de borst te slaan
The Netherlands must stop beating itself on the chest
b. Hunze wil zichzelf niet al te zeer op de borst kloppen
Hunze doesn't want to knock itself on the chest too much
c. Ze verloren zichzelf soms in tactische varianten
They lost themselves in tactical variants
d. Hij verbeeldt zichzelf oogcontact te hebben
He imagines himself to have eye contact

The idiomatic expression *zich/zichzelf op de borst kloppen* (*to boast*) occurs with a strong reflexive 47 times (30% of the time). A few other idiomatic expressions behave similarly. One explanation might be that the idiomatic readings are still transparently linked to the non-idiomatic, accidental reflexive, reading, leading to a certain amount of interference between the two uses.

verb	nonrefl		refl		*zich*		*zichzelf*	
	#	%	#	%	#	%	#	%
straf (*to punish*)	1060	95.7	47	4.3	2	4.2	45	95.8
bescherm (*to protect*)	4921	96.4	186	7.6	95	51.1	91	48.9
vastketenen (*to chain*)	24	34.8	45	65.2	43	95.6	2	4.4

Table 1: Counts and percentages for nonreflexive and reflexive use, and use of weak and strong reflexive pronouns.

5 Statistical Analysis

We used linear regression to determine to what extent there is a correlation between reflexive use of a (non-inherent reflexive) verb and the relative preference for a weak or strong reflexive pronoun.

The data we are dealing with has the form shown in table 1. Establishing a correlation between the percentage of nonreflexive use and the percentage of occurrences of the strong reflexive *zichzelf* with the verb is problematic because the distribution of the percentage of nonreflexive use is far from normal. This is illustrated in figure 1 (left), which shows the percentages in sorted order.[3] A better alternative is to use the ratio of nonreflexive over reflexive use, and the ratio of strong reflexive use over weak reflexive use, and take the log values of these. For nonreflexive use, this gives the distribution in the right pane of figure 1, which is more evenly spread out.

As before, we limit our analysis to verbs that occur at least 10 times with a reflexive meaning and at least 50 times in total, distinguishing uses by subcategorization frames. Figure 2 (left pane) plots the ratio of nonreflexive use over reflexive use (x-axis) against the ratio of strong reflexive forms over weak reflexive forms (y-axis) for all objects. Linear regression (shown as the solid line in fig. 2) gives an r^2 correlation coefficient of 0.162 (statistically significant at $p < 0.001$), with a standard error of 2.07. This means that the ratio of nonreflexive over reflexive use accounts for 16% of the variance in the ratio of strong reflexive over weak reflexive use.

If we count as non-reflexive uses only cases where a verb occurs with a pronoun (as suggested by Haspelmath), 594 verbs remain with frequencies above the cut-offs we used. Linear regression over this data set gives an r^2 of 0.293, and a slightly lower standard error (1.98). If we only consider third person personal pronouns

[3] Statistical analysis was done with R (`http://cran.at.r-project.org`), following the techniques described in Baayen (2008).

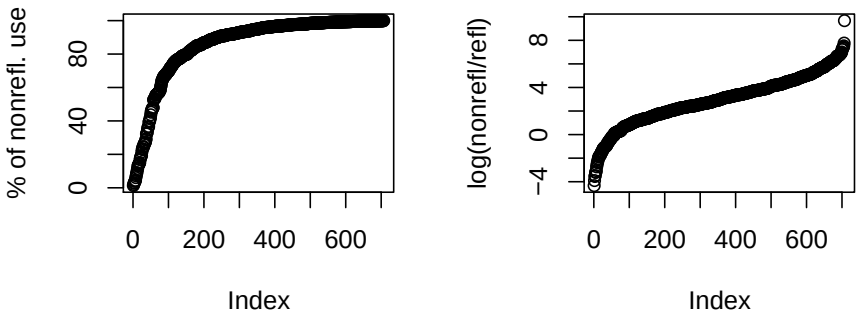

Figure 1: Distribution of percentage of nonreflexive use and ratio of nonreflexive over reflexive use

only (*hem (him), haar (her), hen (them)* and *ze (them)*), 500 verbs remain. We now obtain the result given in fig. 2 (right pane), with an r^2 of 0.332 and a standard error of 1.97.

These results are in line with the findings in Hendriks, Spenader, and Smits (2008). They also observed that restricting object counts to personal pronouns gives a better result than counting all NP-objects. However, for the 32 verbs for which they collected data, they obtain an r^2 of 0.456. As we obtain an r^2 of 0.332, the question arises what might explain this difference. We extracted all verbs from the data-set for personal pronouns that were also used in Hendriks, Spenader, and Smits (2008). 24 of these verbs were sufficiently frequent in our data-set. Linear regression over this limited set gives an r^2 of 0.547 and a standard error of 1.7. One reason for the higher score (compared to Hendriks *et al.*) might be the fact that we take subcategorization frames into account. Another reason might be our use of different frequency cut-offs. What the result also shows, is that our method of data collection in itself does not introduce more noise than the method in Hendriks, Spenader, and Smits (2008). The fact that we obtain a lower score on the larger set of verbs could be due to the fact that the 32 verbs used by Hendriks, Spenader, and Smits (2008) were collected from examples used in the literature. Apparently, these verbs are particularly suitable for demonstrating the statistical correlation to be investigated. Once one takes the full set of verbs into account, however, a fair number of outliers are added as well.

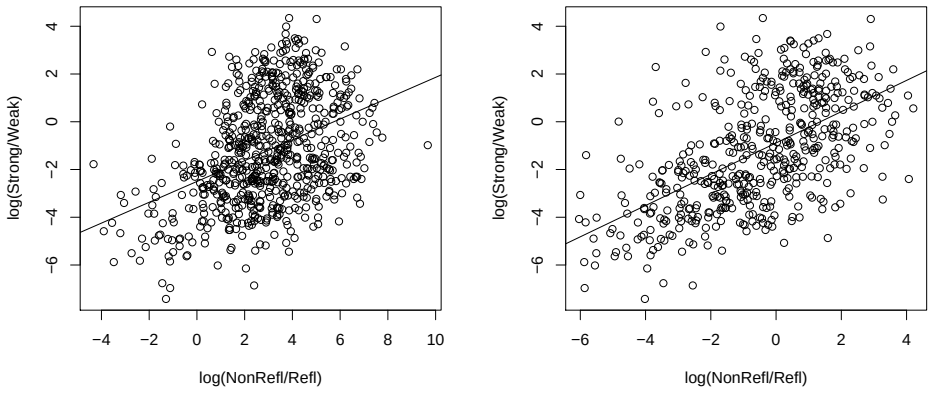

Figure 2: Nonreflexive vs reflexive use compared with strong reflexive over weak reflexive use counting all NP-objects (left) and counting only pronouns (right).

6 Discussion

One of the major ways in which this work tries to improve upon earlier work is by using more data, looking at more verbs (hundreds rather than 30-50) and by using better data (by distinguishing verbs by their subcategorization frames). The assumption is that more data will lead to a better model, and will compensate for irregularities introduced by the fully automated process. Looking at more data did lead to higher correlations for each of the data collection methods, though this effect is not distinguishable from the effect of separating verbs by subcategorization frame.

But looking at more verbs did not give higher correlations. The highest correlation was obtained with the verbs studied by Hendriks, Spenader, and Smits (2008). These are verbs that routinely appear in the literature as good examples of accidental reflexives. One explanation is that these verbs are relatively frequent (although not necessarily frequent in our corpus), and that frequent verbs are the ones for which a speaker may have an expectation of self-directedness or other-directedness. Another explanation is that these verbs in particular might have relatively few different senses, or that they are overwhelmingly used with a sense that has the potential to be both self- or other-directed.

It is still not clear why the ratio of pronominal objects to reflexive objects predicts so much better than taking all objects into account. There are two possible explanations. First, it may be that this restriction in a way also filters out uses

	zichzelf	*zich*		*zichzelf*	*zich*
alleen (only)	109	1	*nu (now)*	16	1
ook (also)	214	9	*wel (certainly)*	14	0
niet (not)	30	9	*min of meer (more or less)*	21	0
slechts (only)	2	0	*alleen maar (only)*	13	1
zelfs (even)	7	0	*zo (that way)*	12	0

Table 2: Choice of reflexive immediately following focus particles

of verbs with senses that essentially cannot be used reflexively. By only counting pronominal objects as non-reflexive objects, the sense of the verb has to be one where the action can be performed on another agent. This would lead to more accurate data (though less data) and may be responsible for the better results.

The other explanation comes from theoretical syntax, Principle A and B of the Binding Theory (Chomsky, 1981) suggests that personal pronouns and reflexives are in complementary distribution when the subject and the object are both animate. In other words, there is a potential for reflexive action only in the case of an animate subject. This means that the ratio for a given action to be self- or other-directed is only reliable if we limited our counts to cases where the subject and object are both animate.

Strictly speaking, comparing the ratio of pronominal objects to reflexive objects doesn't actually give us the ratio of self- vs. other-directed events. This is because we also potentially count cases where the subject is inanimate and the object is a personal pronoun. However, the few corpus studies of grammatical role and animacy that have been done show that the combination of an inanimate subjects with an animate objects is disprefered. Bouma (2008) gives results for spoken Dutch with data for 2,345 sentences from the *Corpus Gesproken Nederlands*. 243 of the sentences had animate objects but among these only 8 (or 3%) occurred with an inanimate subject. Using data from written texts, Øvrelid (2004) looked at 1,000 randomly sampled sentences from the Oslo corpus of Norwegian. 98 of the 1,000 sentences studied had animate objects and of these only 24 had an inanimate subject (24%).

Still, we are able to account for between 30-53% of the data (depending on what dataset is used) using only one predictive factor: how frequently the verb is used with a reflexive object. However, it is also clear that other factors play a role in choosing between a strong and reflexive form. Only strong reflexives can be coordinated, fronted and phonetically focused. This suggests we should take such additional factors into account as well. But coordination of reflexives is rare, and focus or phonetic stress is hard to determine automatically. In a limited number

of cases, one might try to determine focus by taking the preceding expression into account. If the word preceding the reflexive object is a focusing particle, we expect the reflexive following to be *zichzelf*. Table 2 shows that this is indeed the case for a number of expressions that associate with focus.

Factors such as position in the sentence could also be checked. For example, we expect only strong reflexives to be fronted, so we would expect more strong reflexives in initial sentence position. Further, because only strong reflexives can receive sentential accent we would also expect strong reflexives to occur sentence finally more often than weak reflexives (with accidental reflexive verbs). It would be interesting to collect data for the (relative) sentence position of the reflexive (i.e. distance (in words or constituents) from the governing verb or end of the sentence), and to investigate whether a correlation can be found between position and reflexive choice. Geurts (2004) suggests yet another factor. Even non-reflexive verbs like *toedienen* (*to inject oneself*) can use *zich* if the context makes clear the action is a habitual event. This suggests that the presence of temporal adverbs indicating frequency could also play a role. If we can find methods to collect the relevant data automatically, it would be interesting to incorporate them in a multivariate analysis in future work.

Acknowledgements

Jennifer Spenader's work was supported by grant 016.064.062 from the Netherlands Organisation for Scientific Research (NWO).

References

Baayen, R.H. 2008. *Analyzing Linguistic Data*. Cambridge University Press.

Bouma, Gerlof 2008. Starting a Sentence in Dutch. A corpus study of subject- and object-fronting. Groningen Dissertations in Linguistics, 66.

Bouma, Gosse and Geert Kloosterman. 2007. Mining syntactically annotated corpora using xquery. In Branimir Boguraev and Nancy Ide et al., editors, *Proceedings of the Linguistic Annotation Workshop (ACL 07)*, Prague.

Chomsky, Noam 1981. Lectures on Government and Binding. Foris, Dordrecht

Geurts, Bart. 2004. Weak and strong reflexives in dutch. In *Proceedings of the ESSLLI workshop on semantic approaches to binding theory*, Nancy, France.

Haspelmath, Martin. 2004. A frequentist explanation of some universals of reflexive marking. Draft of a paper presented at the Workshop on Reciprocals and Reflexives, Berlin.

Hendriks, Petra, Jennifer Spenader, and Erik-Jan Smits. 2008. Frequency-based constraints on reflexive forms in dutch. In *Proceedings of the 5th International Workshop on Constraints and Language Processing*, pages 33–47, Roskilde, Denmark.

Ordelman, Roeland, Franciska de Jong, Arjan van Hessen, and Hendri Hondorp. 2007. Twnc: a multifaceted Dutch news corpus. *ELRA Newsletter*, 12(3/4):4–7.

Øvrelid, Lilja. 2004. Disambiguation of syntactic functions in Norwegian: modeling variation in word order interpretations conditioned by animacy and definiteness. In Fred Karlsson, editor, *Proceedings of the 20th Scandinavian Conference of Linguistics*, Helsinki.

Reinhart, Tanya and Eric Reuland. 1993. Reflexivity. *Linguistic Inquiry*, 24:656–720.

Smits, Erik-Jan, Petra Hendriks, and Jennifer Spenader. 2007. Using very large parsed corpora and judgement data to classify verb reflexivity. In Antonio Branco, editor, *Anaphora: Analysis, Algorithms and Applications*, pages 77–93, Berlin. Springer.

van Noord, Gertjan. 2006. At last parsing is now operational. In Piet Mertens, Cedrick Fairon, Anne Dister, and Patrick Watrin, editors, *TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*. pages 20–42.