Using Very Large Parsed Corpora and Judgment Data to Classify Verb Reflexivity

Erik-Jan Smits[1], Petra Hendriks[1], and Jennifer Spenader[2]

[1] Center for Language and Cognition
[2] Artificial Intelligence
University of Groningen
Dutch Department, Faculty of Arts, P.O. Box 716 - 9700 AS Groningen,
The Netherlands
E.J.Smits@rug.nl

Abstract. Dutch has two reflexive pronouns, *zich* and *zichzelf*. When is each one used? This question has been debated in the literature on binding theory, reflexives and anaphora resolution. Partial solutions have attempted to use syntactic binding domains, semantic features and pragmatic concepts such as focus to predict reflexive choice, but until now no experimental data either in favor of or against one of these theories is available. In this paper we look at reflexive choice on the basis of empirical data: a large scale corpus study and an online questionnaire. On the basis of the results of both experiments, we are able to predict the choice between the two reflexive items in Dutch without assuming a distinction between verbs that occur with *zich* or *zichzelf* a priori (cf. a distinction in terms like 'inherent reflexivity' (Reinhart and Reuland, 1993)). Instead, we examine the distribution of *zich* and *zichzelf* using the Clef corpus, a 70 million word Very Large Corpus of Dutch. The corpus is tagged and parsed. This allows us to identify the typical action the verbs are used to describe: reflexive or non-reflexive actions. Regression analysis shows that, by doing so, 21% of the distribution of the two reflexive items in Dutch can be predicted. Using the verb reflexivity found in the corpus study even allows us to explain 83% of the participants' choices in the online study between *zich* and *zichzelf*. As such, both the corpus study and the online questionnaire confirm the group of verbs called 'inherent reflexive verbs' without postulating the group beforehand. We further discovered that even inherently reflexive verbs, which are argued to never co-occur with *zichzelf*, sometimes had *zichzelf* chosen as the preferred argument in the questionnaire, and to a lesser degree, in the corpus suggesting that the verb classes are tendential and not categorical.

1 Two Reflexives, One Meaning?

Dutch, like German, French, Swedish and Danish, but unlike English, has two reflexive pronouns: *zich* and *zichzelf*, both unspecified for gender, number and case:

A. Branco (Ed.): DAARC 2007, LNAI 4410, pp. 77–93, 2007.
© Springer-Verlag Berlin Heidelberg 2007

(1) Jan wast zich/zichzelf.
Jan washes SE/SELF
'Jan washes himself'

(2) Jan schaamt zich/*zichzelf.
Jan schames SE/*SELF
'Jan is ashamed of himself'

(1) can be used with both *zich* and *zichzelf*, while (2) seems only to be possible with *zich*. There has been much theoretical debate about what features predict the choice of *zich* or *zichzelf*. The choice has been argued to be the result of syntactic constraints (Broekhuis 2004, Reuland and Koster 1991), to be strongly affected by semantic properties of the verb (Haeseryn et al. 2002 (Algemene Nederlandse Spraakkunst, ANS), Reinhart and Reuland 1993, Lidz 2001) by the degree of affectiveness of the situation (Everaert 1986, Geurts 2004), or by the placement of focus (Everaert, 1986). However, as far as we know there are no large-scale corpus studies or questionnaire studies documenting the use of *zich* and *zichzelf*. Such data, however, is important for several reasons: first, heuristics for the types of objects a given verb tends to co-occur with can improve parsing. Second, the choice of reflexive *zich* with a non-reflexive verb is suggested to be related to the habitualness of the event in the context. Confirming this empirically would mean we have a new surface clue to habitual events, an interesting result for natural language understanding. Third, the acquisition of reflexives and pronouns is a major topic in child language. To correctly make materials and interpret results for Dutch and other language with two reflexives we need to know what their uses are. Finally, the results should be relevant to the choice of the reflexive in natural language generation.

The purpose of this study is to see to what degree a large-scale corpus study and an online questionnaire can help predict the choice between *zich* and *zichzelf*. Through an analysis of the distribution of *zich* and *zichzelf* among predicate types, we also address the existence of a number of different classes of reflexivity which can be found in the literature (among other terms *inherent reflexive verbs*, *necessarily reflexive verbs*, *accidental reflexive verbs*). We do this by examining the use of each predicate and looking at how often the action denoted by the verb is reflexively performed in the corpus compared to how often it is performed to some other party. The experimental data points out that it is only possible to do so if both reflexive and non-reflexive transitive uses are taken into account, considering both corpus and questionnaire data.

2 *Zich* vs. *Zichzelf*

In Binding Theory in Generative Grammar approaches to syntax, Principle A governs the use of reflexive forms:

Principle A: A reflexive must be bound in its local domain.[1]

Because the distinction between the use of pronouns and reflexives can largely be explained purely on the basis of syntactic criteria (i.e. their binding relations), a similar syntactic based approach has been suggested for explaining the distribution of the two reflexive forms in Dutch (*zich* and *zichzelf*) (Broekhuis 2004, Reuland and Koster 1991). Reinhart and Reuland (1993) however argue against standard Binding Theory and its characterization in terms of the syntactic characteristics of the NP, assuming instead a much closer relation between anaphora and argument structure. Put differently, they claim that reflexivity is a syntactic property of predicates. Most important for the current paper, they make a syntactic distinction between *zich* and *zichzelf*, respectively called SE and SELF anaphora.

The debate in the literature seems to have focused on two questions: 1) Are there different classes of verbs that differ in their choice of a SE or SELF reflexive argument? and 2) Is there a difference in meaning between a SE and a SELF reflexive?[2] A third question that has yet to be consistently addressed is 3) What effect does context have on the felicity of a SE or SELF anaphor?

Looking first at the question of verb classes, most theorists claim that there are at least two classes: inherently reflexive verbs and regular transitive verbs. The Dutch Grammar ANS (Haeseryn et al, 2002) identifies a group of verbs as "noodzakelijk reflexieve werkwoorden", or necessarily reflexive verbs, including such verbs as *zich vergissen* (to err), or *zich zorgen maken* (to worry). These verbs are claimed to only occur with *zich* and never with *zichzelf*, and further can never occur with a non-reflexive object. They also recognize "toevallige reflexieve werkwoorden", or accidental reflexive verbs, which can occur with *zich*, *zichzelf*

[1] For current purposes, it suffices to take 'local domain' as the sentence containing the reflexive and, taking a standard Chomskyan perspective, define 'binding' as a relation between two elements A and B for which holds that A and B are co-indexed and A c-commands B (i.e. there is a structural relation between A and B or, more precisely, A c-commands B or A is in a higher structural position than B). Crucially, pronouns must be free in their local domain, i.e. are not co-indexed with an element A in the same sentence and are not c-commanded by this A.

[2] We know that there are several differences between the two forms related to information structure. Only *zichzelf* can occur in coordination and focus positions such as questions answers, in clefts or in topicalization, e.g. (i) and (ii) based on examples from Geurts (2004):

(i) De trainer heeft *zich/zichzelf en zijn hond aangemeld.
The trainer has *SE/SELF and his dog registered
'The trainer registered himself and his dog'

(ii) De kok heeft alleen *ZICH/ZICHZELF gesneden.
The cook has only *SE/SELF cut
'The cook only cut himself'

Zich can never be phonetically focused, and has been argued to not be possible in a context in which there are no salient alternatives (see Reinhart 2003 and Geurts 2004).

and with a third person object. These two groups should be distinguished from a third group of verbs that can occur with a third person object or *zichzelf* but never with *zich*. The groups themselves are defined according to the types of arguments that are felicitous and as such cannot be used to predict the use of *zich* and *zichzelf.* ANS has nothing to say about a possible semantic difference between SE and SELF reflexives.

In summary, based solely on their distribution with different arguments ANS distinguish between three types of verbs.

(3) Necessarily reflexive verbs (only with *zich*)

a. Jan vergist zich
Jan mistake SE
'Jan makes a mistake'

b. *Jan vergist zichzelf
Jan mistake SELF
'Jan makes a mistake'

c. *Jan vergist de hond
Jan mistake the dog
'Jan makes the dog a mistake'

(4) Non-reflexive verbs (never with *zich*)

a. *Zij begrijpen zich niet
They understand SE not
'They don't understand themselves'

b. Zij begrijpen zichzelf niet
They understand SELF not
'They don't understand themselves'

c. Zij begrijpen de melkboer niet
They understand the dairy farmer not
'They don't understand the dairy farmer'

(5) Accidental reflexive verbs (with both *zich* and *zichzelf*)

a. Jan wast zich
Jan washes SE
'Jan washes himself'

b. Jan wast zichzelf
Jan washes SELF
'Jan washes himself'

c. Jan wast de melkboer
Jan washes the dairy farmer
'Jan washes the dairy farmer'

Unlike ANS, Reinhart and Reuland (1993) recognize only two classes of predicates: inherently reflexive predicates[3], that can occur with *zich*, and

[3] As Reinhart and Reuland 1993) point out, the observation that some verbs express an intrinsic reflexive relation between its arguments actually goes back to Jesperson (1933), Gleason (1965) and Partee and Bach (1981) (where it is attributed to Montague).

transitive predicates. Similar to ANS they offer no independent criteria to determine whether or not a given verb belongs to one or the other class. This is only determinable by looking at which arguments the verb can co-occur with. Reinhard and Reuland derive the third class, accidental reflexive verbs, by arguing that *zichzelf* is an operator that can reflexivize a transitive predicate,[4] imposing an identity relation on the two arguments of a predicate:

> “a transitive predicate that is not inherently reflexive may turn into a reflexive predicate if reflexivity is marked on one of its arguments with a SELF-anaphora” (1993: 662)

Zich is not an operator since it can only occur with a predicate that is already inherently reflexive. Thus when the same surface verb form occurs with *zich* it is the necessarily reflexive predicate form, and when it occurs with *zichzelf* it is the transitive predicate that has been turned into a reflexive predicate through the operator *zichzelf.* According to Reinhart and Reuland, this explains why *zich* is not allowed in (6), but is in (7); *wassen* (to wash) has an inherently reflexive and a transitive lexical entry; *haten* (to hate) has no inherently reflexive predicate counterpart, and its transitive entry can only be turned into a reflexive predicate by using a SELF anaphor. Since this is not the case in (6), *zich* is ungrammatical.

(6) Jan haat *zich/zichzelf
Jan hates self
‘Jan hates himself’

(7) Jan schaamt zich/*zichzelf
Jan is ashamed of self
‘Jan is ashamed of himself’

In sum, Reinhart and Reuland state that it is the type of predicate that determines the distribution of *zich* and *zichzelf* since inherently reflexive predicates will be able to just use *zich.* For accidental reflexive verbs that allow both *zich* and *zichzelf* like (1), they must postulate an ambiguity in the lexicon: the same surface verb *wassen* (to wash) has both an inherently reflexive and a transitive form.

Lidz (2001), like Reinhard and Reuland, believes that there is both a transitive and an inherently reflexive lexical entry for accidental reflexive verbs like, e.g. *scheren* (to shave). But Lidz considers *zich* and *zichzelf* to have different meanings. Reinhart and Reuland’s account predicts that all reflexively marked predicates correspond to the same type of semantic reflexivity, regardless of how the reflexivity was achieved, i.e. lexically on the verb or with a SELF operator. Lidz (2001) argues against this conclusion, by convincingly showing that there is in fact a difference. Consider this example: Ringo Starr goes into Madame Tussaud’s wax museum. Once he sees his own statue, he notes that they have portrayed him with a beard. But he does not have a beard in real life. Displeased,

[4] See for a slightly different view Keenan (1988), who argues that the SELF anaphor turns a transitive verb into an intransitive one.

Ringo decides to shave the beard off the statue. According to Lidz, it is felicitous in Dutch to utter (9), but not (8) (from Lidz, 2001:128).

(8) Ringo scheert zich
Ringo shaves SE
'Ringo shaves himself'

(9) Ringo scheert zichzelf
Ringo shaves SELF
'Ringo shaves himzelf'

Conversely, if Ringo does have a beard in real life and he decides that he looks better without one given the way they portrayed him in the Madam Tussaud's wax museum (i.e. in this case without a beard) and begins to shave his own face, (8) is felicitous and (9) is not. Both sentences are marked for reflexivity, in terms of Reinhart and Reuland either lexically on the verb or syntactically with *zichzelf*, yet they differ in the situations in which they are true or false. The operation of changing the transitive *scheren* (to shave) into a reflexive action by applying a *zichzelf* operator results in a reflexive shaving action that differs in meaning from the inherently reflexive lexical entry *scheren*.

In order to capture this observation, Lidz (2001) replaces the distinction between SE and SELF anaphora of Reinhart and Reuland with what he considers a more semantic one: he calls SE reflexive modified predicates *pure-reflexive predicates* and SELF reflexive modified predicates *near-reflexive predicates*. In near reflexive predicates the reflexive *zichzelf* object is a function of the subject, but not identical with it, unlike in the true reflexive predicates:

(10) Semantic/pure reflexive predicates:
$\lambda x\,[P(x,x)]$

(11) Near-reflexive predicates:
$\lambda x\,[P(x,f(x))]$

Lidz's (2001) account gives different semantic representations for (8) and (9) cases with reference only to the sentence itself. However it's clear that in his example the context plays an important role in determining which form is felicitous, and he gives no account of how to distinguish contexts appropriate for *zich* from those appropriate for *zichzelf*. Since the inherent reflexive form and the transitive form of the verbs are homophones, and there is currently no way to determine when we are dealing with the inherently reflexive form or the transitive form besides looking at the argument, this account also cannot help us predict the choice of *zich* vs. *zichzelf*.

We need to find independently motivated features that correlate with the choice between *zich* and *zichzelf*. Work in this direction is found in Zubizarreta (1987), who looks at the *semantic affectiveness* of the predicate. The idea of affectiveness was originally discussed in Anderson (1979). A verb is +affective if its action results in a change in the abstract or physical state of its object. Because +affective verbs cause a change, the events they refer to are also necessarily delimited, though the converse is not true: not all verbs that refer to delimited events are +affective. We can call a verb +affective if it denotes a delimited event favoring a coreferential interpretation between the subject and object, as a result of which actions such as *brushing*, *washing* or *shaving* are +affective while *admiring* or *promising* are -affective.

By using affectiveness Zubizarreta tries to distinguish between inherent reflexive verbs and transitive verbs without defining them in terms of the potential to use *zich* or *zichzelf.*[5] Zubizarreta (1987) begins by arguing that +affective verbs have an internal argument. She gives *eat* as a typical example. When used intransitively *eat* is argued to have a hidden argument of food. She presents *out*-prefixation as a distinguishing test; verbs with affected internal arguments can take *out*-prefixation, e.g. John outate Bill, while standard transitive verbs like *confuse* can't, e.g. *John outconfused Bill. Zubizarreta (1987) claims that inherently reflexive predicates are a subset of verbs that have affected internal arguments, and supports the claim with data from Dutch based on an observation in Everaert (1986). Everaert argues that *zich* can't be a co-argument with the subject in its binding domain because it behaves as a clitic and is a bound anaphor, illustrating this with examples like the following, where (12) illustrates that the *zich* cannot be a co-argument with *zij*, and (13) where it can function as an argument in an adjectival small clause. (examples from Everaert 1986:126):

(12) *Zij begrijpen zich niet.
They understand themselves not.
'They do not understand themselves'

(13) Marie maakt zich niet druk
Marie makes herself not stressed
'Marie is not stressed'

In (13) *zich* is not considered by Everaert to be a co-argument of *druk maken* (to stress out), but in (12) it is considered to be a co-argument of *begrijpen* (to understand). But there are a number of clear counterexamples to this claim. Actually for accidental reflexive verbs *zich* can be a co-argument with the subject:

(14) Jan wast zich.
Jan washes himself
'Jan washes himself'

(15) Jan verbergt zich
Jan hides himself
'Jan hides himself'

Zubizarreta (1987) explains the binding of *zich* with the subject in (14) and (15) by stating that these verbs are 'inherently reflexive'. She then argues that the verbs are syntactically *intransitive*, despite their misleading appearance of transitivity; this is because *zich* can be considered to be an internal affected argument. Further, Zubizarreta (1987) claims these verbs are semantically *transitive*. The

[5] Note that Zubizarreta is concerned with the difference between transitive verbs and inherent reflexive verbs and actually her paper never mentions *zichzelf.*

fact that they realize *zich*, which mistakenly appears to be a subject-coargument, is merely because they wear their semantics on their sleeve. (14) and (15) are thus not counterexamples to the generalization that *zich* cannot be syntactically bound by a subject because *zich* isn't a syntactic argument. *Zich is* a semantic argument because the verbs are +affective and their meaning requires that they act on some sort of object.

Zubizarreta's account means verbs like (14) and (15) should pattern with verbs like (13) and not like (12). A potential problem is that within the group of accidentally reflexive verbs like (14) and (15), some verbs seem to require *zich*, e.g. *verbergen* (to hide), while others like *wassen* (to wash) realize *zich* optionally, e.g. intransitively. Further, Zubizarreta (1987) claims that the realization of *zich* with the last group is a lexical idiosyncracy.

Zubizarreta classifies verbs differently than ANS, Reinhart and Reuland and Lidz: the accidental reflexive verbs with their reflexive uses are classified together with the inherently reflexive verbs.

By not taking *zichzelf* into consideration Zubizarreta misses an important characteristic that distinguishes the class of verbs like (13) from those like (14) and (15). The first group never co-occurs with *zichzelf*, while the latter group can. Also, the first group cannot take a third person argument, while the second group clearly can and does. There is also clearly a tangible difference in meaning between 'Jan wast' (Jan washes) and 'Jan wast zich' (Jan washes himself); in the former Jan can be washing any object but in the latter he must be washing himself. Finally, it is hard to think of what evidence could confirm Zubizarreta's assumption that there is a hidden semantic argument in certain inherently reflexive verbs like *wassen* (to wash) when they are used with *zich*.

Affectiveness has also been appealed to by Jakubowicz (1992) to explain the binding possibilities of the Danish SE reflexive *sig*, quite similar to Dutch *zich*. Jakubowicz argues that verbs that allow local binding with the Danish SE reflexive *sig* are only those that are +affective. Because these verbs also co-occur with the Danish SELF reflexive *sig selv* and with non-reflexive objects, the class of +affective verbs seems to coincide with the class of accidentally reflexive verbs. The local binding ability in *sig* is attributed to an argument present in the +affective verbs, again because the action predicated by the verb must act on something or someone, and thus is concrete enough to be bound locally. In contrast to Zubizarreta (1987) Jakubowicz considers the +affective verbs to be syntactically, as well as semantically, transitive.

Zubizarreta's and Jakubowicz's work is interesting in that they try to ground the idea of reflexivity in terms unrelated to the features they are trying to predict. However, because the definition of +affectiveness is quite vague, it doesn't help us that much with predicting the choice between *zich* and *zichzelf*; we lack a method for objectively determining affectiveness.[6] Because the above explanations are either circular or incomplete we will work with the surface characteristics considering there to be three classes of verbs.

[6] How general the process of out-prefixation is, isn't clear; further, it isn't applicable to Dutch.

2.1 Flexible Class Membership?

A question that has not been addressed in the earlier work is whether membership in one of the above three classes is categorical or whether membership is flexible. Geurts (2004) brings up an interesting example of a case where explicitly emphasizing the reflexivity of an event can make *zich* possible with a verb that most informants would in a neutral context immediately classify as non-reflexive (Geurts, 2004: 4).

(16) De zuster dient *zich/zichzelf opium toe
The sister injected *SE/SELF opium in
'The nurse injected herself with opium'

(17) Betty dient zich/zichzelf weer eens opium toe
Betty injected SE/SELF once again opium in
'Betty injected herself once again with opium'

A nurse normally gives medicine to patients, i.e. others. However, if we know that Betty is a drug addict, and habitually injects herself with opium, when we refer to a token event of this type, *zich* becomes possible. It seems then that the class which the verb falls into is changed by purely pragmatic characteristics, e.g. pragmatic coercion. This example is problematic for the classification of Reinhart and Reuland (1993) and Lidz (2001) because a verb that is generally *not* consider to have a inherently reflexive version seems to acquire just such a lexical entry when the context is appropriate. This suggests that the choice of argument is a regular alteration more than the existence of two lexical entries.

Inherent reflexivity as a semantic feature is perhaps evaluated against the sum of all events in our experience, in which case normal injections are not reflexive. But it seems that in a delimited context, e.g. Betty's life, the sum of all events can be the realm of evaluation, in which case injecting is typically a reflexive event, and *zich* becomes possible.

The example given by Lidz (2001) could also be analyzed along these lines. The use of either *zich* or *zichzelf* to express verb reflexivity results in a difference in meaning because of the habitualness of the situation; in the Madame Tussaud examples, *zich* is possible when Ringo shaves his own face (e.g. not the statue's) because that is a normal reflexive shaving action. Because Ringo's shaving his statue is not standard shaving, *zichzelf* is preferred.

2.2 Towards Objective Predictors

The problem of the current classification of verbs seems to be that they are all based on the feature we want to predict: inherent reflexive verbs are defined as those verbs that only occur with *zich*. A verb is non-reflexive if it cannot occur with *zich*. Also, the divisions that exist of what verbs fall into each category have been determined entirely introspectively. It is therefore not clear to what degree they are correct, and to what degree they have been subjectively determined by the analyst. Further, for the accidental reflexive verbs where both *zich* and *zichzelf* are possible it would be advantageous to know if one was more frequent than the other, and under what conditions each occurs. There seems to be a relationship between the frequency with which an action is performed reflexively and the 'class' to which the verb belongs.

To gain more objective information about the use of reflexive arguments we decided to do two empirical studies, a corpus study and an online questionnaire. We predict that for verbs that frequently occur with a third person object, and therefore are referring to a non-reflexive event, the use of *zichzelf* will be more frequent than the use of *zich* among the accidental reflexive verbs. Further, we predict that verbs that are seldom used to refer to non-reflexive events will have a higher frequency of co-occurrence with *zich* than with *zichzelf.* Further we also predict that because argument co-occurrence has to do with the ratio of the frequency of reflexive or non-reflexive actions in the world, the classes are not lexically determined.

Our aims are, first, to experimentally verify the difference between necessarily and accidental reflexive verbs, and, second, to experimentally test the hypothesis that the choice between *zich* and *zichzelf* correlates with the typical relation a predicates denotes with respect to its argument(s). The theoretical literature mentioned above predicts that we will not find any necessarily reflexive predicates with *zichzelf.* This follows from the definition of *zichzelf* as an operator which turns a non-reflexive predicate into a reflexive predicate; *zichzelf* can only be applied to a non-reflexive predicate and not to a necessarily reflexive predicate. Conversely, the theories predict that non-reflexive predicates typically occur with *zichzelf* and not with *zich.* In order to answer these questions we did two empirical studies.

3 Empirical Data

3.1 Corpus Study: Method and Results

For the corpus study we used the CLEF (Cross-Language Evaluation Forum) corpus for Dutch made up of 72 million words and 4,150,858 sentences taken from the full content of the 1994 and 1995 Dutch daily newspapers of Algemeen Dagblad and NRC Handelsblad (Jijkoun, Mishne, and de Rijke 2003). The corpus was parsed with the LFG-based Alpino parser (Bouma, van Noord, and Malouf 2001).

We focused on 60 verbs, where 28 of the verbs were defined as inherently reflexive by ANS (Haeseryn et al. 2002). Third person subjects with objects were searched for in the corpus for these 60 verbs. We counted how often each verb occurred with a reflexive *zich*, *zichzelf* or with a non-reflexive object.

First a comparison of *zich* and *zichzelf* was made. The results are displayed in the boxplots in figure 1 in which necessarily reflexive verbs and accidental reflexive verbs are plotted on the x-axis and the use of *zich* on the y-axis (in percentages of the total number of transitive usages). Statistical analysis shows that the distribution of *zich* and *zichzelf* in corpus data to a great extent confirms ANS's classification. *Zich* is significantly more often found to occur with the verbs that are labeled necessarily reflexive verbs in ANS than with the accidental reflexive verbs. A t-test shows that *zich* is significantly more often used with necessarily reflexive verbs (mean = 82.4%, sd. = 25.5, std. error mean = 4.6) than with accidental reflexive verbs (mean = 99.3%, sd. = 2.7, std. error mean = 4,5) (t(58) = -3,5, p = .001). Most members of the class of necessarily reflexive verbs never occur with *zichzelf*. One of a few exceptions is *ontpop* (turn into), that was used 638 times with *zich*; however, it was used once with *zichzelf*:

(18) Aan het slot van zijn tweede informateurschap heeft Tjeenk Willink zichzelf ontpopt als het activistische type.
'At the end of his second informator-ship Tjeenk Willink turned (himself) into the activist type.'

Because *zich* is also possible with (18), and because this is not a typical focus position, it is not clear how to distinguish this usage. A verb like *straffen* (to punish) is, in line with our predictions, seldom found to occur with *zich* (cf. fig. 1 in which *straffen* is marked as an outlier with an asterix). Below is one of the two examples *straffen* did occur with a SE-anaphor, which interestingly enough, is also an example with *straffen* and a SELF anaphor, where the SELF anaphor is probably chosen for contrast:

(19) Straft de tragische held Oedipus *zich* lijfelijk, deze Eddie
Punish the tragic hero Oedipus SE physically, this Eddie
straft *zichzelf* door het onmogelijke te willen ...
punishes SELF through the impossible to want ...
'While the tragic hero Oedipus punishes himself physically, this Eddie punishes himself by wanting the impossible ... '

For current purposes, an even more important question for the corpus study is: can we predict the distribution of *zich* and *zichzelf* without *a priori* assuming a distinction à la ANS? Or, put differently, can we find a relationship between the frequency with which a verb occurs with a reflexive object versus a non-reflexive object, and the frequency with which the same verb in only reflexive events occurs with *zich* versus *zichzelf*. For this reason we looked at all transitive uses of each verb, including uses with a non-reflexive object. We made a simple linear regression analysis using the use of *zich* and the frequency of reflexive usages as regressors. The regression analysis shows that 21% of the use of *zich* can be predicted by the frequency of reflexive events with the same verb (R^2 = 0.21%, t(63) = 3.9, p < .001).

We can explain 21% of the data by knowing how frequently a verb occurs with a reflexive action from the corpus data alone. However, people might use *zich* or *zichzelf* for other reasons. To see how closely the corpus data reflects the intuitions of naïve speakers, we also did an online questionnaire. Further, because many of the verbs occurred infrequently even in our 70 million word corpus, we felt it was important to supplement results based on a handful of examples with intuitions.

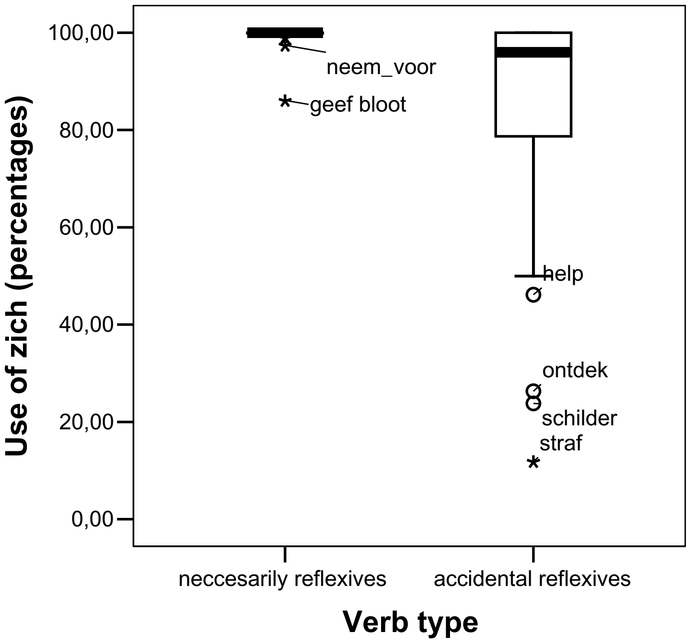

Fig. 1. The use of *zich* (expressed on the y-axis in terms of the percentage of the total number of arguments a predicate is found to occur with) for both necessarily reflexive verbs and accidental reflexive verbs (following the definitions of ANS (Haeseryn et al. 2002)). Translations of the displayed verbs are as follows: *neem voor*,'have intentions' *geef bloot*, 'reveal', *help*, 'help', *ontdek*, 'discover', *schilder*, 'paint', *straf*, 'punish'.

3.2 Online Questionnaire: Method and Results

Twenty-nine adult native speakers of Dutch took part in an online test where they were asked to make a forced choice between *zich* and *zichzelf* as the best argument for 78[7] potentially reflexive verbs. The stimuli were presented in short sentences like (20):

(20) Maria schaamde ________
'Maria was ashamed of ________'

Similar to the corpus data, the data from the online questionnaire reveals a significant difference between the distribution of *zich* and *zichzelf* for necessarily reflexive and accidental reflexive verbs. *Zich* is used in 21.9% of the cases (sd. = 8.0, std.error mean = 1.5) with accidental reflexive verbs and in 93.7% of the

[7] In the corpus study we excluded a number of otherwise interesting verbs because of the existence of a homonym with a very different sense that would have required checking examples by hand. For example, the verb *wegscheren* can mean 'to shave away' but also, with *zich* in the combination *zich wegscheren* 'get out of here', where only the former is truly transitive.

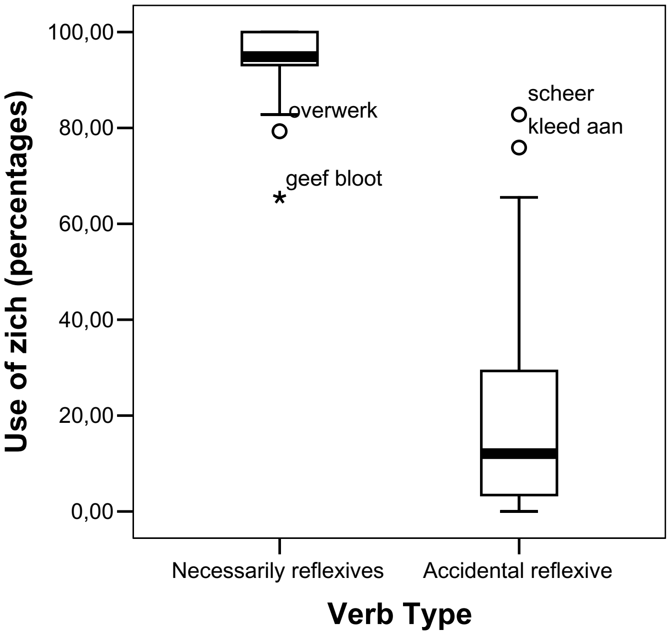

Fig. 2. The use of *zich* in the questionnaire (expressed on the y-axis in terms of the percentage of the total number of arguments a predicate is found to occur with) for both necessarily reflexive verbs and accidental reflexive verbs (following the definitions of ANS (Haeseryn et al. 2002)). Translations of the displayed verbs are as follows: *overwerk*,'overwork' *geef bloot*, 'reveal', *scheer*, 'shave' and *kleed aan*, 'dress'.

cases (sd. = 26.5, std.error mean = 4.7) with necessarily reflexive verbs, $t(58)$ = -13.8, p < .001). This again experimentally confirms Haeseryn et al.'s distinction between necessarily and accidental reflexive verbs.

3.3 Comparing the Data

Comparing the results from the questionnaire study with the results from the corpus study, statistical analysis reveals a significant difference between the use of *zich* for Haeseryn et al.'s necessarily and accidental reflexive verbs (respectively 93.7% versus 99.3% for neccessary reflexive verbs, $t(27)$ = -4.5, p < .001, and 21.9% versus 82.4% for accidental reflexive verbs, $t(31)$ = -11.2, p < .001). We suspect that the difference has to do with the sparse data problem for some of the reflexive verbs in the corpus, which was the motivation for doing the questionnaire study in the first place. Since the difference between the two classes is still the same we see both types of data as complementary confirmation of Haeseryn's classification.

To test the hypothesis that the distribution of *zich* and *zichzelf* in the questionnaire also correlates with verb reflexivity, we compared the online choices for *zich* or *zichzelf* in the questionnaire for the 60 verbs that occurred in both

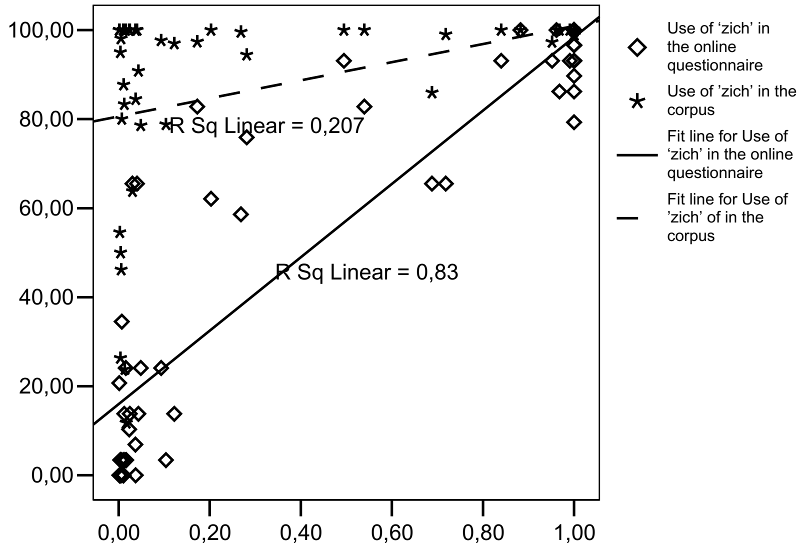

Fig. 3. Typical reflexive usage of each verb expressed on the x-axis in terms of the percentage of reflexive uses among all its uses (i.e. reflexive and non-reflexive transitive uses) versus the use of *zich* per verb on the y-axis. The dotted line represents the correlation between the use of *zich* in the corpus study and the typical reflexive usage of each verb (found in the corpus study). The solid line represents the correlation between the use of *zich* found in the online questionnaire and the typical reflexive character of each verb as found in the corpus study.

experiments (see Appendix A). Simple linear regression shows that 83% of the distribution is predicted along these lines (R^2=0.83, $t(61)$ = 16.9, p < .001). This shows that the inherent reflexive nature of the verb, defined as the frequency with which a verb is used to refer to a reflexive action or a non-reflexive transitive action in the corpus, is a correct predictor of the distribution of *zich* and *zichzelf* in the online questionnaire.

As a next step, we used Fisher's r to z-test to statistically analyze the difference between the correlation coefficients found for the distribution of *zich* and *zichzelf* in the corpus data and the questionnaire with respect to the typical predicate structure. This revealed a significant difference between the correlation in the corpus study and the online questionnaire regarding the reflexive nature of the verb and the distribution of *zich* and *zichzelf* (z = -5.5, p < .001). Put differently, the data from the corpus study gives us a better picture how often a verb is used with a reflexive or a non-reflexive action. This in turn significantly improves our ability to model the use of *zich* and *zichzelf* in the questionnaire. This leads us to conclude that the distribution of *zich* and *zichzelf* can be predicted solely on the basis of corpus and judgment data. No distinction has to be made a priori between necessarily and accidental reflexive predicates along the lines of ANS.

4 Discussion

The results show that it is possible to predict to a large degree what class a verb belongs to (either to the class of necessarily reflexive verbs or to the class of accidental reflexive verbs). Moreover, we have shown that a combination of a corpus study and an online questionnaire allows us to do so. Several reasons motivate the decision to supplement the corpus work with judgment data. Even with a corpus of over 70 million words, it is not possible to find reflexive uses of all the verbs that can possibly occur with a reflexive meaning. For example, *ruiken* (to smell) only occurred with reflexive objects twice in the corpus (however 451 times with a non-reflexive object). This is a very small number to draw conclusions from. Moreover, the corpus data alone was not a perfect predictor for the distribution of *zich* and *zichzelf*, we also needed to looked at judgment data. For example, the verb *schamen* (to be ashamed, reflexive in Dutch) was only used once with *zichzelf* in the corpus, and the case was a direct translation from English to Dutch which may have influenced the choice. However, in the online test 6 respondents preferred *zichzelf* instead of *zich*. The context sentence was extremely short and neutral, so the preference for *zich* might be explained by the tendency for *zich* to avoid focus positions. The end of the sentences is a typical focus position in Dutch. However, this explanation would make it hard to explain why the other 23 respondents preferred *zich*.

Because we did the corpus work for the most part automatically, some of the results might be incorrect. Incorrect parses of imperative and topicalized sentences were found when the data was hand-checked. This has certainly introduced some noise in the data, but it is just this type of parsing error that empirically founded reflexive classes might be able to help avoid.

Do corpus results and judgment results give us a way to predict the choice of reflexive? Yes and no. We can derive the main classes without assuming a priori classes, but we cannot predict individual choices for accidental reflexive verbs. We can use the corpus results to confirm what verbs belong to the class of inherent reflexive verbs (preference for *zich*), and to confirm what verbs are typically 'non-reflexive' (preference for *zichzelf*). But because the online study shows that subjects can deviate from the predictions, other factors such as focus and the habitualness of the action need to be considered. Various participants in the online questionnaire pointed in a similar way at the habitualness that seems to play a role in the meaning of *Het meisje snijdt zich* 'The girl cuts herself' for unintentional cutting, versus *Het meisje snijdt zichzelf* for intentional cutting.

Zubizarreta's work brought up a possible additional factor. She suggests that the intransitive uses of the accidental reflexives somehow play a role in the frequency of the use of *zich* vs. *zichzelf*. This could be tested. If the choice to omit *zich* is totally idiosyncratic, then we should be able to count purely syntactically intransitive uses of e.g. *wassen* (to wash) as being reflexive uses. We can then consider whether ratios of reflexive uses to non-reflexive uses are more predictive if we count syntactically intransitive uses as being among the reflexive transitive uses. We leave this for future work.

Revisiting the questions brought up at the beginning of the paper, we found evidence confirming the existence of at least three classes of reflexives, though the membership is not completely categorical as many researchers have thought, with the categories being just strong tendencies. We were not able to evaluate whether or not there is a difference in meaning between SE and SELF reflexives because we did not look at individual examples, but we think the fact that we found exceptions among the class of necessarily reflexive verbs that took *zichzelf* as an argument seems to suggest that there is some difference: how otherwise can we explain this deviation from the strong trend for these verbs only to occur with *zich*? Finally we did not address the question of how context effects the choice in our experimental studies because this also involves manual checking of examples, but note that this is an obvious future endeavor.

In sum, we have found evidence that verbs do roughly belong to classes of necessarily reflexive verbs and accidental reflexive verbs. We conclude that the corpus data alone does not completely predict the choice between *zich* and *zichzelf.* Because judgment data reveals significantly different patterns we conclude that both sets of data are necessarily to make a good model. By doing so, we are able to, unlike previous research, predict class membership to a high degree based on the frequency with which the verb is used to refer to a reflexive action or a non-reflexive action. In doing so we come to the conclusion that the transitive uses of the verb and the reflexive uses are actually related. In fact, it strongly calls into question the underlying assumption in the work of Reinhart and Reuland (1993), Lidz (2001) and Zubizarreta (1987) that there are two identical surface forms mapping to two different underlying verb forms, the inherently reflexive predicate form and the transitive predicate form. Remember, the motivation for this distinction was to be able to account for the difference between verbs that have no transitive form and allow only *zich* and those that allow both. This analysis requires postulating two distinct lexical entries for each accidental reflexive verb surface form. Since these authors offer no independently motivated way to prove two distinct forms exist, and we can distinguish them on the basis of the frequency of all the arguments they co-occur with (e.g. non-reflexive as well as reflexive), it seems unnecessary to maintain this view.

References

Anderson M. (1979). Noun phrase structure. Unpublished doctoral dissertation, MIT, Cambridge, MA.

Bouma, G., G. van Noord, and R. Malouf. (2001). Alpino: Wide-coverage computational analysis of Dutch. In Computational Linguistics in The Netherlands 2000. Rodopi

Broekhuis, H. (2004). The referential properties of noun phrases I (2nd edition). Modern grammar of Dutch occasional papers 1, University of Tilburg.

Everaert, M. (1987). *The syntax of reflexivization.* Foris Publications, Dordrecht, The Netherlands / Riverton, USA

Geurts, B. (2004). Weak and Strong Reflexives in Dutch, In: Philippe Schlenker and Ed Keenan (eds.), Proceedings of the ESSLLI workshop on semantic approaches to binding theory.

Gleason, H. (1965). Linguistics and English Grammar. New York: Holt, Rinehart and Winston.

Haeseryn, W., K. Romijn, G. Geerts, J. de Rooij & M.C. van den Toorn.(2002). Algemene Nederlandse Spraakkunst. Second, totally revised version of 2002. Groningen/Deurne, Martinus Nijhoff /Wolters Plantyn

Jakubowicz, C. (1992). Sig en danois: syntaxe et acquisition, in H.-G. Obenauer and A. Zribi-Hertz (eds.), Structure de la phrase et thêorie du liage, Presses Universitaires de Vincennes, Saint Denis, pp. 121-149.

Jesperson, P. (1933). Essentials of English grammar. London: Allen and Unwin (1983).

Jijkoun, V., Mishne, G. & M. de Rijke. (2003). Preprocessing Documents to Answer Dutch Questions. In: Proceedings 15th Belgian-Dutch Conference on Artificial Intelligence (BNAIC'03), 2003.

Keenan, E. (1988). On semantics and the binding theory. In J. Edwards (ed) *Explain language universals*. John Hawkings. Oxford: Blackwell.

Lekakou, Marika. (2005). Reflexives in contexts of reduced valency: German vs. Dutch. In The Function of Function Words and Functional Categories, Dikken, M. den and C. M. Tortora (eds.). John Benjamins.

Lidz, J. (2001). Condition R. Linguistic Inquiry **32** (1), 123-140

Partee, B. and Bach, E. (1981). Quantification, pronouns, and VP-anaphora. In Formal methods in the study of language. Mathematisch Centrum, Amsterdam University.

Reinhart, T. & Reuland, E. (1993). Reflexivity. Linguistic Inquiry **24**, 656 - 720

Reuland, E. and J. Koster. (1991). Long-distance anaphora: an overview. In Long-distance anaphora, ed. J. Koster and E. Reuland, 1-25. Cambridge:Cambridge University Press

Zubizarreta, M.L. (1987) *Levels of Representation in the Lexicon and in the Syntax.* Foris Publications: Dordrecht, the Netherlands / Providence RI, USA

Appendix

Necessarily reflexive verbs tested (28), following ANS (Haeseryn et al. 2002):

abonneren (to subscribe to), *bedrinken* (to get drunk), *inbeelden* (to imagine), *behelpen* (to make do), *beijveren* (to), *bemoeien* (to interfere with), *beraden* (to think over), *beroemen* (to boast), *indenken* (to image)), *distantiëren* (to dissociate), *gedragen* (to behave), *bloot geven*, *generen* (to feel embaressed), *schuilhouden* (to hide), *inleven* (to imagine), *misdragen* (to misbehave), *voornemen* (to resolve), *ontfermen* (to take pity on), *ontpoppen* (to turn out to be), *ontspinnen* (to lead to), *overwerken* (to overwork), *schamen* (to be ashamed of), *vergrijpen* (to attack)), *verhouden* (to be in proportion), *verkneukelen* (to chuckle) , *verloven* (to engage), *verslikken* (to choke) , *verspreken* (to make a slip of the tongue).

Accidental reflexive verbs tested (32):

aaien (to pet), *achtervolgen* (to follow), *bedekken* (to cover), *beschermen* (to protect), *bewonderen* (to admire), *bijten* (to bite), *binden* (to bind), *geven* (to give), *ingraven* (to bury), *helpen* (to help), *horen* (to hear), *kietelen* (to tickle), *aankleden* (to dress), *knippen* (to cut), *kussen* (to kiss), *lachen* (to laugh), *opmaken* (to make up), *omhelzen* (to hug), *ontdekken* (to discover), *prikken* (to prick), *ruiken* (to smell), *scheren* (to shave), *schilderen* (to paint), *schoppen* (to kick), *slaan* (to hit), *snijden* (to cut), *straffen* (to punish), *tekenen* (to draw), *tillen* (to lift), *verstoppen* (to hide), *vertellen* (to tell), *zien* (to see).