Preposition stranding and using R-pronouns for humans in Dutch

Bachelor thesis

Anne Kingma
s0823376

29-08-2012

Helen de Hoop

Table of Contents

1. Introduction

One of the best things about songs and poems is that they extend grammar rules. Things that would be weird or even completely unacceptable in an everyday conversation, are suddenly considered beautiful and poetic. The sentence in (1) is taken from a song (*Genoeg* by Eefje de Visser). The second line consists of four PPs, which seem very similar on first sight - but one of them does not fit. In Dutch, a preposition cannot take a third person pronoun like *dit* "this". Normally, the pronoun would turn into an R-pronoun (Van Riemsdijk, 1978) and move to before the preposition: *hiervan*, literally "hereof", meaning "of this". That of course does not go very well with this song, though. For a few more examples, see (2).

(1) Ik heb nooit genoeg
I have never enough
van jou van mij van ons van dit
of you of me of us of this
"I've never had enough of you, of me, of us, of this"

(2)

a.	op	het	dak	*op	het	er-op
	on	the	roof	on	it	there-on
	"on the roof"					"on it"
b.	in	dat	huis	*in	dat	daar-in
	in	that	house	in	that	there-in
	"in that house"					"in that"
c.	van	welke	fiets	*van	wat	waar-van
	of	which	bike	of	what	where-of
	"of which bike"					"of what"

This process is called *preposition stranding*. It is of course very common in English as well, except in English, it does not involve R-pronouns. An R-pronoun is a locative pronoun, but in this construction it is not interpreted as such.
Both the movement and the pronoun turning into an R-pronoun are obligatory. Doing neither leaves the PP ungrammatical as is shown in (2), although in specific contexts, it can be acceptable (like in the song in (1)). It is possible to have something that looks like an R-pronoun follow the preposition, like in (3) (PP in boldface):

(3) De bus rijdt **tot** **daar.**
the bus drives **till** **there**
"The bus goes up to there."

Here, however, the pronoun does refer to an actual location, it is used with its original locative meaning. As Helmantel (2002) puts it, it is used deictically. Only then is it possible to leave it after the preposition.
The fourth logical possibility, moving the pronoun but not turning it into an R-pronoun, is ungrammatical as well (example (4a)). The same goes for full DPs (4b), which do not have an R-version to turn into. It should probably be noted, however, that some speakers do use these

sentences in spoken language - Dutch seems to be moving towards English in this respect (see Schippers 2012).

(4) a. ***Dat** heb ik geen tijd **voor.**
that have I no time **for**
Intended meaning: "I don't have time for that."
b. ??**Dat** **kind** heeft ze **mee** gespeeld.
that **child** has she **with** played
Intended meaning: "She played with that child."

There is one other type of word that allows preposition stranding: Quantificational pronouns. In this case, however, the stranding is optional, which makes them fall between full NPs and other pronouns. Still, preposition stranding requires the pronoun to become an R-pronoun. (5) shows the possibilities for *alles/overal* "everything/everywhere", it is the same for *niets/nergens* "nothing/nowhere" and *iets/ergens* "something/somewhere".

(5) a. Ik heb **aan** **alles** gedacht.
I have **on** **everything** thought
b. Ik heb **overal** **aan** gedacht.
I have **everywhere** **on** thought
c. *Ik heb **alles** **aan** gedacht.
I have **everything** **on** thought
d. *Ik heb **aan** **overal** gedacht.
I have **on** **everywhere** thought
(Intended) meaning for all four: "I thought of everything."

There is one more restriction on preposition stranding which is important for this thesis: It is not allowed when the pronoun has a human referent. This is shown in the sentences in (6). The marked sentences are not ungrammatical per se, but they do not correspond to the intended meaning (they are interpreted with a non-human referent).

(6) a. Ik vraag het **aan** **hem / #daar-aan.**
I ask it **to** **him /** **there-to**
"I'm asking him."
b. **Met** **wie** **/** **# Waar-mee** werkt zij samen?
with **whom** **/** **where-with** works she together?
"Who does she collaborate with?"
c. Wij hebben **voor** **niemand** **/** **# nergens** **voor** gezorgd.
we have **for** **nobody** **/** **nowhere** **for** taken.care
"We didn't take care of anyone."

Lestrade et al. (2010) link this phenomenon to Kuryłowicz (1964) who suggests that the dative and locative cases originate as variants of the same case, with the dative being meant for animates and the locative for inanimates. Aristar (1996) shows some cases of languages where the dative and locative cases have a complementary pattern based on animacy: The locative case is unmarked with

an inanimate noun and marked with an animate noun; the dative case is marked with an inanimate noun and unmarked with an animate noun. Apparently, locative forms do not like to be combined with animate referents (Lestrade et al., 2010). As mentioned above, an R-word is a locative pronoun. The incompatability of an R-word with a human referent could be the reason why preposition stranding is not allowed with humans.
There seems to be more to it, however. The sentences below, for example, seem fine:

(7) **Daar** heb ik altijd ruzie **mee.**
there have I always fight **with**
"I'm always fighting with him/her!"

(8) het meisje **waar-mee** ik danste
the girl **where-with** I danced
"the girl I danced with"

Even though they are prescriptively considered bad style and discouraged in written language (for example by the ANS (Haeseryn et al., 1997)), they are common in spoken language. A more descriptive grammar by Broekhuis (2002) states: "Despite the fact normative grammars generally state that of (23a) and (23b), only the former is acceptable, constructions like (23b) are the ones normally found in colloquial speech." (p. 263) His (23) is repeated here as (9):

(9) a. de jongen op wie ik wacht
the boy for whom I wait
'the boy I am waiting for'
a'. *de jongen wie ik op wacht
b. (?)de jongen waarop ik wacht
the boy where-for I wait
'the boy I am waiting for'
b'. de jongen *waar* ik *op* wacht

Broekhuis (2002), p. 263, example (23)

Lestrade et al. (2010) do mention these sentences, but they do not provide an explanation for them. Apparently, something is more important here than the restriction of not using R-pronouns for humans. Note, however, that the version *without* preposition stranding is grammatical as well in these cases, and does not change the meaning of the sentence. This suggests some kind of competition between these two forms.

There are two constraints working against each other. One constraint makes the pronoun move and turn into an R-pronoun and leave its preposition behind. Why this happens is not a part of this thesis, but see for example Lestrade et al. (2010) who relate it to scrambling. On the other hand, there is the resistance of human nouns to be referred to with an R-form which was discussed above. In a basic sentence with unmarked word order, like (6a), this is enough to completely rule out preposition stranding. Evidently, however, this is not always the case. Either the reason for the pronoun to move is stronger in these cases, the resistance to a locative form for humans is weaker, or both.

A type of sentence in which referring to humans with an R-pronoun is known to occur is a relative clause (cf. Broekhuis (2002), Haeseryn et al. (1997)). However, the Dutch Language Union (*Nederlandse Taalunie*) advises people to avoid it and children in schools are taught not to use it. The reason why it occurs more often with relative clauses probably lies in the force that is making the pronoun move. In this case it is not scrambling, it is wh-fronting, which is obligatory in Dutch. There is simply no way of *not* moving it:

(10) a. dat ik **met** **het** **meisje** danste
that I **with** **the** **girl** danced
"…that I danced with the girl"

b. *het meisje ik **met** **wie** danste
the girl I **with** **whom** danced
Intended meaning: "The girl with whom I danced"

c. het meisje **waar** ik **mee** danste
the girl **where** I **with** danced
"The girl I danced with"

d. het meisje **met** **wie** ik danste
the girl **with** **whom** I danced
"The girl with whom I danced"

The alternative to stranding the preposition is moving it along with the pronoun (see (10d)), which is called *pied-piping*. Apparently in some cases, this is less favourable than referring to a human with a locative pronoun. But when and why do people choose either form? A lot of factors could influence this choice. A simple answer would be 'speaker variation' (Lestrade et al., 2010; De Vries, 2006) – things like age, region and education probably do play a part, but for the purpose of this thesis I will mostly ignore them. Another factor could be the relative clause itself (see for example Johansson & Geisler (1998) or Hoffman (2005) for English, or De Vries (2006) who also includes German and Dutch).
In this thesis, however, I will focus on the referent of the R-pronoun. I hypothesize that some referents resist association with an R-form more strongly than others, and some therefore force the preposition to pied-pipe and some don't . This causes (at least part of) the variation in referring to humans with R-pronouns.

I will use the Spoken Dutch Corpus (CGN, see section 2) to investigate this phenomenon. The first question I will answer is how common it is in spoken language to refer to human referents using an R-pronoun compared to using a regular pronoun? Secondly, I will look at the referents themselves: Are there any properties of the referent that determine its likelihood to be referred to with an R-pronoun, and if so, which? I will start by briefly introducing the CGN in section 2. In section 3, I will explain the methods I used in searching the Corpus. The results will be analyzed in section 4 and discussed in section 5, after which the conclusion will follow in section 6.

2. The Spoken Dutch Corpus

The Spoken Dutch Corpus (*Corpus Gesproken Nederlands*, CGN, 2006) consists of about 9 million words spoken by native speakers of Dutch in the Netherlands and Flanders, recorded between 1998 and 2003. The first complete version of the corpus was published in 2004; version 2, which I have used for this thesis, was published in 2006. About 3,3 million words were produced by Flemish speakers, 5,6 million are from The Netherlands. The corpus is organized into several components, each containing a different type of speech (see Table 1).

Table 1. Overview of the data in the CGN, divided into components. (CGN, 2006)
(VL = data originating from Flanders; NL = data originating from The Netherlands)

Component		Total	VL	NL
a.	Spontaneous conversations ('face-to-face')	2,626,172	878,383	1,747,789
b.	Interviews with teachers of Dutch	565,433	315,554	249,879
c.	Spontaneous telephone dialogues (recorded via a switchboard)	1,232,636	489,1	743,537
d.	Spontaneous telephone dialogues (recorded on MD with local interface)	853,371	343,167	510,204
e.	Simulated business negotiations	136,461	0	136,461
f.	Interviews/discussions/debates (broadcast)	790,269	250,708	539,561
g.	(political) Discussions/debates/ meetings (non-broadcast)	360,328	138,819	221,509
h.	Lessons recorded in the classroom	405,409	105,436	299,973
i.	Live (e.g. sports) commentaries (broadcast)	208,399	78,022	130,377
j.	Newsreports/reportages (broadcast)	186,072	95,206	90,866
k.	News (broadcast)	368,153	82,855	285,298
l.	Commentaries/columns/reviews (broadcast)	145,553	65,386	80,167
m.	Ceremonious speeches/sermons	18,075	12,510	5,565
n.	Lectures/seminars	140,901	79,067	61,834
o.	Read speech	903,043	351,419	551,624
Total		8,940,098	3,285,631	5,654,644

The whole corpus has been transcribed orthographically, lemmatised and part-of-speech-tagged. In addition to that, phonetic transcription, syntactic annotation and/or prosodic annotation is available for part of the data. For this thesis, the part-of-speech-tagging and the syntactic annotation are the most important. I will therefore elaborate on them below.

2.1 Part-of-Speech-tagging in the CGN

The data was first tagged automatically, then checked and if necessary corrected manually. The tag set consists of 316 tags and is based on the ANS (*Algemene Nederlandse Spraakkunst*, Haeseryn et al., 1997). A detailed description of the tag set and the general guidelines used in tagging the corpus can be found in Van Eynde (2004, in Dutch).

In tagging prepositions, prepositions that preceded their complement received a different tag from prepositions that followed their complement (postpositions). This was of course very useful in my research. For pronouns, on the other hand, no distinction was made between regular demonstrative

pronouns and R-pronouns (which are the same morphologically). Because of that, I needed the syntactic annotation in order to find preposition stranding constructions.
Something else I needed to take into account was the fact that when the R-pronoun and preposition end up next to each other in a sentence, they are written as one word:

(11) a. het meisje **waar** ik **mee** danste
the girl **where** I **with** danced
b. het meisje **waarmee** ik danste
the girl **where.with** I danced
"The girl I danced with"

In the CGN, each word receives one and only one tag. Words like these don't have their own tag. They are tagged as adverbs and are therefore hard to find using only POS-tagging. For these reasons I decided to use the syntactic annotation as well, even though this is available for only part of the corpus.

2.2 Syntactic annotation in the CGN

About one million words in the Spoken Dutch Corpus were annotated syntactically. This is the part of the corpus that was used for this thesis. For an overview of this data and to which components they belong, see Table 2.

Table 2. Overview of the data for which a syntactic annotation is available. (CGN, 2006)
(VL = data originating from Flanders; NL = data originating from The Netherlands)

Component		Total		
			VL	NL
a.	Spontaneous conversations ('face-to-face')	447,113	146,745	300,368
b.	Interviews with teachers of Dutch	59,751	34,064	25,687
c.	Spontaneous telephone dialogues (recorded via a switchboard)	89,819	19,886	69,933
d.	Spontaneous telephone dialogues (recorded on MD via a local interface)	6,257	6,257	0
e.	Simulated business negotiations	25,485	0	25,485
f.	Interviews/discussions/debates (broadcast)	100,250	25,144	75,106
g.	(political) Discussions/debates/meetings (non-broadcast)	34,126	9,009	25,117
h.	Lessons recorded in the classroom	36,064	10,103	25,961
i.	Live (e.g. sports) commentaries (broadcast)	35,116	10,13	24,986
j.	Newsreports/reportages (broadcast)	32,744	7,679	25,065
k.	News (broadcast)	32,689	7,305	25,384
l.	Commentaries/columns/reviews (broadcast)	32,502	7,431	25,071
m.	Ceremonious speeches/sermons	7,077	1,893	5,184
n.	Lectures/seminars	23,056	8,143	14,913
o.	Read speech	44,144	44,144	0
Total		1,006,193	337,933	668,260

Details on the choices made in the syntactic annotation can be found in Hoekstra et al. (2003). The goal of the annotation was to keep it as simple as possible while providing as much information as possible for different kinds of users. The result is a fairly theory-neutral kind of dependency structure. Basically, it uses phrase nodes only when that is the most straightforward way of annotating a structure. Whenever they don't have direct practical use, they are dropped. It does, for example, include NPs, but no VPs.

3. Method

To answer the research questions, I used two types of constructions which contain an R-pronoun with a preposition. The first are relative constructions, like the sentence in (8) (here repeated as (12)):

(12) het meisje **waar-mee** ik danste
the girl **where-with** I danced
"The girl I danced with"

As mentioned in the introduction, however, it is prescriptively considered bad style to refer to a human with an R-pronoun. It is therefore likely that some people actively try to avoid it and use the construction with a regular PP more often than they would otherwise. This, in addition to the consideration that using only relative constructions would not provide much data, was a reason for me to add another construction to my research. These are sentences like in (7) (here repeated as (13)):

(13) **daar** heb ik altijd ruzie **mee**
there have I always fight **with**
"I'm always fighting with him/her!"

Using demonstrative R-pronouns for humans is common as well, but it has received much less attention than relative R-pronouns. The average native speaker is hardly aware of it. Self-correction effects should therefore be minimal. The decision to use topicalised sentences is based on Bouma's (2008) research. He used the CGN to determine the properties of elements in the so-called Vorfeld in Dutch. The Vorfeld in short is the part of the sentence in Dutch and German before the finite verb, where amongst other things topicalised elements belong. The prescribed alternative to the sentence in (13) is topicalising the whole PP:

(14) **met** **hem** heb ik altijd ruzie
with **him** have I always fight
"I'm always fighting with him!"

According to Bouma's (2008) findings, there are two things somewhat unfortunate about this construction. Firstly, there is a PP in the Vorfeld, which is not very common compared to preposition stranding. Secondly, personal pronouns (like *hem* 'him') don't like being in the Vorfeld, as opposed to demonstrative pronouns (like the R-pronoun *daar*). Because of that, preposition stranding seems most likely to occur in combination with topicalisation: Even though referring to a human being with an R-pronoun is not ideal, the alternative has problems too, which might outweigh the humanness constraint.

3.1 Finding the sentences in the CGN

For both of these types (relative clauses and topicalised PPs) I searched for three kinds of constructions: With the R-pronoun connected to the preposition (the examples in (a)), with the R-

pronoun disconnected from the preposition (the (b) sentences) and with a regular PP without an R-pronoun (the (c) sentences).

(15) a. het meisje **waarmee** ik danste
the girl **where.with** I danced
"The girl I danced with"
b. het meisje **waar** ik **mee** danste
the girl **where** I **with** danced
"The girl I danced with"
c. het meisje **met** **wie** ik danste
the girl **with** **who** I danced
"The girl with whom I danced"

(16) a. **daarmee** heb ik altijd ruzie
there.with have I always fight
"I'm always fighting with him/her!"
b. **daar** heb ik altijd ruzie **mee**
there have I always fight **with**
"I'm always fighting with him/her!"
c. **met** **hem** heb ik altijd ruzie
with **him** have I always fight
"I'm always fighting with him!"

Below I will explain for each of these types of sentences what criteria I used to find them in the corpus. The exact syntax of the searches can be found in appendix A. In most cases, these strategies returned some results that were irrelevant for the present thesis, in addition to the sentences I was looking for. While manually annotating the data (see section 3.2), these were removed from the data set.

3.1.1 Relative clauses

Relative clauses are tagged as such in the syntactic annotation of the CGN, they are therefore not hard to find. The relative pronoun or the phrase that contains it is tagged as 'head', the rest of the clause is 'body'.

In order to find phrases like (15a) (R-pronoun and preposition attached), I used the POS-tagging. As mentioned in section 2, these words are tagged as adverbs. Without the syntactic annotation, this makes them hard to find, but restricting the search to adverbs which are the head of a relative clause returns mostly relevant results.

Constructions like (15b) (R-pronoun and preposition disconnected) were more complicated to find. In these cases, the R-pronoun is the head of the relative clause, so I could use a combination of the POS-tagging and syntactic annotation to find them. The R-pronoun can however function as a regular locative pronoun as well, a distinction which is not made in the POS-tagging. I therefore limited the search to relative clauses which contained a PP with a postposition, using the syntactic annotation. (There is a separate POS-tag for postpositions as opposed to prepositions.)

Finally, examples like (15c) (a regular PP without R-pronoun) were easiest to find: In these sentences, the whole PP is annotated as the head of the relative clause. Simply searching for relative clauses with a PP as the head was enough.

3.1.2 Topicalised PPs

Topicalisation is not explicitly tagged in the CGN. I used word order to find topicalised sentences. Dutch (as well as German) is a verb-second language: In main clauses, the finite verb always has to come second. In unmarked sentences, the subject takes the first place, automatically forcing the rest of the sentence to after the finite verb (17a). Only when something is topicalised can it take the first place in a sentence, before the finite verb (17b). The subject then ends up third, directly following the verb. Putting more than one constituent before the finite verb leaves the sentence ungrammatical (17c,d).

(17) a. ik heb hem gezien
I have.1SG him seen
"I saw him"

b. hem heb ik gezien
him have.1SG I seen
"It's him I saw"

c. *ik hem heb gezien
I him have.1SG seen

d. *hem ik heb gezien
him I have.1SG seen

In the CGN, the finite verb is analyzed as the head of the clause. For main clauses, therefore, an R-pronoun that precedes the head of the clause can be considered topicalised. Subordinate clauses don't have verb-second word order, so topicalisation is hard to find. Because of that, I only used main clauses for this thesis.

Sentences like (16a) (R-pronoun and preposition attached) cannot be found using only POS-tagging and syntactic annotation. Searching for topicalised adverbs does of course include these sentences, but along with them a huge amount of other adverbs. Adverbs of time and place, for example, are frequently topicalised. The only way to find them without also retrieving a lot of irrelevant results, is by using the orthographic transcription and searching for each possible word seperately. In order to minimise the chances of missing a possibility, I used the Lexicon tool which comes with the CGN to make a list of all words tagged as preposition, and then took from that list all prepositions that can take an R-pronoun (there were 22, see Appendix B).

There was only one common preposition that was excluded deliberately: *om* 'around'. The reason for this is that the version of this preposition with an R-pronoun (*daarom*) has become lexicalized with the meaning 'because'. Its ending up sentence-initially very often has nothing to do with R-pronouns whatsoever.

For the type of (16b) (R-pronoun and preposition disconnected) the POS-tag was more useful, but like with the relative R-pronoun, the same tag is used for regular demonstrative pronouns. In order to exclude those, I used the syntactic annotation to search only for pronouns that are the complement of a preposition.

In order to find sentences like (16c) (the whole PP topicalised), it is not enough to simply search for topicalised PPs. Unlike with the relative clauses, it is possible here for the complement of the preposition to be first or second person. With the POS-tagging it is possible to add some more detail: there is one tag for second and third person pronouns in oblique case (which is compulsory when following a preposition). The relevant results can then be filtered manually. For the sake of completeness, I also searched for demonstrative pronouns, even though they are generally considered ungrammatical with a preposition (cf. *van dit* 'of this' in (1)).

3.2 Annotation

Animacy is not annotated in the CGN. This had to be added manually. The searches described in section 3.1 yielded a lot of inanimate results which needed to be filtered out. During this process, all irrelevant results caused by imperfections in the search strategies and occasionally by mistakes in the POS-tagging or syntactic annotation were discarded as well.

After that, for each of the selected sentences had to be determined what word or phrase the pronoun referred to exactly. This part was done by two annotators independently and then compared afterwards, to ensure reliability. In addition to that, some properties of the word were annotated that might have an effect on its chances of being referred to with an R-pronoun. Some of these are straightforward features that are known to have an effect on several syntactic phenomena (Hopper & Thompson, 1980; Yamamoto, 1999; De Swart, 2007; Bouma, 2008; Van Bergen & De Swart, 2010). Some others were included specifically because they are some measure of the 'human-ness' of the NP.

Animacy is not always a clear dichotomy, it is a scale (Van Bergen, 2011; Comrie, 1989; Kuno, 1976). Humans are more animate than animals which are in turn more animate than inanimate objects (Comrie, 1989). Kuno (1976) uses the notion of *empathy* to explain this animacy scale: The better we can empathize with it, the higher up the scale it comes. We as humans can empathize best with other humans.

Even within these groups, however, there can be variability. All referents in this thesis are human, still, some can be considered more clearly so than others. A proper name, for example, can make even inanimate objects be treated as animates (Yamamoto, 1999). A doll becomes a person once it has a name. This can be related to Kuno's (1976) empathy as well: Once something has a name, it is easier to be empathized with. Similarly, it is easier to empathize with someone you know personally than with a celebrity you only know from stories. I assume that the higher on the humanness scale the referent is, the more important the humanness constraint is and the more likely it is to not be referred to with an R-form. Considering there is no previous research on this topic specifically, it is not clear which properties of humanness are most important. Therefore, several different variables related to humanness have been included.

For most variables, one or more example sentences from the CGN are added for clarification. Repetitions and slips of the tongue are removed for ease of reading, but apart from that, the utterances are copied literally.

- *Number*: Two categories, singular and plural.
- *Definiteness*: Definiteness is shown to have an effect on many different syntactic phenomena (Hopper & Thompson, 1980; De Swart, 2007; Bouma, 2008; Van Bergen & De Swart, 2010; amongst many others). It consists of two categories, definite and indefinite.

- *Complex*: An NP was considered 'complex' whenever it contained more than simply a determiner, adjective and noun. These were NPs containing relative clauses and PPs, for example.

(18) en *die mensen* *achter de camera* **daar** hoor je
and those people behind the camera **there** hear you
en zie je nooit wat **van**
and see you never something **of**
"…and *those people behind the camera*, you never see them or hear from them."
CGN: fn000458.83

- *Pronoun*: The reference that was chosen for the analysis was the closest instance of a full NP. Pronouns were avoided because they themselves refer to another linguistic entity. Only when there was no full NP available in the context, the pronoun was used. These were mostly indefinite pronouns like 'all', 'somebody' or 'nothing'.

(19) maar ik heb *iemand* in de hal zitten en
but I have someone in the hallway sit and
daar moet ik een aantal zaken **mee** regelen
there have I a couple things with settle
"…but there is *someone* in the hallway I have to settle some things with."
CGN: fn000887.59

- *Name*: It could be argued this is covered by 'definiteness' already. Although names are indeed always definite, they also capture another aspect of humanness: it is a very personal way to refer to someone.

(20) *Mario* **daar** gaat het ook niet goed **mee**
Mario there goes it also not well with
"*Mario* isn't doing well either."
CGN: fv400100.107

- *Celebrity*: A celebrity is a human being, but generally not one the speaker personally knows. It is therefore a more distant and abstract human being - it is more like a phenomenon than a person. In addition to that, the name of a celebrity is sometimes used to refer to something the celebrity is known for (in case of a singer, for example, his or her music – see also Broekhuis 2002).

(21) a. want *Bach* die kon 't gewoon zelf scheppen dus
because Bach that could it just self create so
voor **hem** was 't allemaal heel gewoon
for him was it all very normal
"because *Bach*, he could simply create it himself so to him it was all completely normal"
CGN: fn007235.124

b. bijvoorbeeld *Marc* *Verwilghen* **daar** hoort men ook bijna
for.example Marc Verwilghen there hears one also almost
niets meer **van**
nothing anymore of
"*Marc Verwilghen*, for example, you hardly hear anything about him nowadays"
CGN: fv400542.73

- *Group*: The group variable includes words that strictly speaking refer to a group of humans, but make it a less direct reference. The words *groep* 'group' and *klas* 'grade' (meaning a group of school children) make up half of this variable. Some other words are *familie* 'family', *brandweer* 'firefighters' (in Dutch a more abstract word comparable to 'police'), *zo'n vijfhonderd* 'about five hundred' and *PSV* 'PSV' (a soccer club).

(22) a. en dan K's Choice da 's *de* *enige* *groep* **waar** ik alle
and then K's Choice that's the only group where I all
CD's **van** koop
cd's from buy
"and then there's K's Choice, that's *the only band* of which I buy every cd."
CGN: fv400195.96

b. *familie* **daar** moet je toch iets **voor** over hebben
family there have you PART something for left have
"*Family* is something you should be willing to sacrifice things for."
CGN: fn000363.204

In appendix C an example of a few annotated sentences can be found.

While discussing the annotation, I will keep the two types of sentences (relative clauses and topicalised PPs) separate. This is because the sentences with topicalised PPs were a lot harder to annotate. Relative clauses typically follow the element they refer to directly. For demonstrative pronouns, the reference can be in the same sentence, but in many cases it is further back in the conversation - in a few cases even 20 or 30 utterances back. Sometimes it was not explicitly mentioned at all.

The first thing to be annotated was the NP the pronoun referred to. Of the 73 relative clauses, the annotators differed on 10 sentences. In six of these cases, they had written down different parts of the same NP and in three cases, they had written down different NPs that referred to the same entity in the conversation. Only one item was actually different.
Of the 137 topicalised PP sentences, the annotations differed on 50 items. Of these, 13 were different parts of the same NP, 14 were different references to the same non-linguistic entity and 23 were actual disagreements.
For the final annotation, it was decided to use the last mention of the referee that was not a pronoun (because a pronoun always refers to something else) and to use the whole of the NP (including for example relative clauses). The items that were really different were checked again, 10 of them were changed, 14 were kept the same as the first version of the annotation.

The second part of the annotation consists of the characteristics of the NP it refers to. These are all catogorical values, therefore the agreement between the two annotators can easily be tested using Cohen's Kappa. In Table 3 an overview of these results can be found, including the interpretation of the values as given by Landis & Koch (1977). All values are significant at the 0.01 level.

Table 3: Agreement of the two annotations measured with Cohen's Kappa (all significant at the .01 level) with an interpretation of the numbers as according to Landis & Koch (1977).

Topicalized PPs			**Relative clauses**		
Category	**Cohen's Kappa**	**Categorization**	**Category**	**Cohen's Kappa**	**Categorization**
Number	0.63	substantial	Number	0.91	almost perfect
Definiteness	0.60	moderate	Definiteness	0.69	substantial
Pronoun	0.85	almost perfect	Pronoun	0.90	almost perfect
Complex	0.46	moderate	Complex	0.68	substantial
Name	0.86	almost perfect	Name	0.82	almost perfect
Celebrity	0.57	moderate	Celebrity	0.27	fair
Group	0.64	substantial	Group	0.64	substantial

As expected, the overall agreement is higher for relative clauses than it is for the topicalized PPs. This is due to the fact that they are harder to annotate: Over half of the differences here are a result of the annotators' having used a different NP. For the variable 'Complex' and to some extent 'Celebrity' and 'Group' some miscommunication is also to blame. These last two however proved to be simply hard to annotate. All cases that differed were checked again for the final annotation.

4. Results

4.1 Initial numbers

As explained in section 3, there are two constructions used for this thesis, and for each construction I searched for three types of sentences. This makes six groups. Below is an example sentence for each group from the CGN.

(23) a. en dus Jacco had nog steeds ook die boer **waarbij** ie
and so Jacco had PART still also that farmer **where-at** he
kon vliegen
could fly
"…and so Jacco still had that farmer as well where he could fly."

CGN: fn000818.40

b. Trio
Trio
waar Eelko een bedrijfje **mee** heeft
where Eelko a company **with** has
"Trio, whom Eelko owns a company with."

CGN: fn000434.208-209

c. d'r zijn wel mensen **bij** **wie** 't mooi staat
there are PART people **at** **whom** it pretty stands
"There are people on which it looks good."

CGN: fn000667.87

(24) a. **daarachter** zit Franck Vanderbroecke
there-behind sits Franck Vanderbroecke
"Behind him is Franck Vanderbroecke."

CGN: fn007162.3

b. en z'n vriendin was een blonde **daar** heb 'k ook
and his girlfriend was a blonde **there** have I also
nog koffie **mee** gedronken nog
PART coffee **with** drunk PART
"…and his girlfriend was a blonde, I had a coffee with her too."

CGN: fn000948.165

c. dus **voor** **hem** was dat een keerpunt in zijn leven
so **for** **him** was that a turning-point in his life
"So for him, that was a turning point in his life."

CGN: fn009128.78

The results for each group are shown in Table 4. For both types of sentences, the regular PPs (which are prescriptively the only correct form) are less than half of the data, although this difference is not significant for the relative clauses (X^2 (1, n = 73) = 1.66, p > .05). For the topicalised PPs, however, it is significant on the .01 level (X^2 (1, n = 137) = 16.12, p < .01).

Table 4: All results of PPs referring to humans in relative clauses and in main clauses where they are topicalised.

	Relative clause		**Topicalised PP**	
	number	**percentage**	**number**	**percentage**
With R-pronoun	42	57,5	92	67,2
Without R-pronoun	31	42,5	45	32,8
Total	**73**	**100**	**137**	**100**

In table 5, the results with an R-pronoun are shown in some more detail, showing whether or not the R-pronoun is attached to the preposition. Again, the difference between the two groups is not significant for the relative clauses (X^2 (1, n = 42) = 0.10, p > .05), but it is for the topicalised PPs (X^2 (1, n = 92) = 62,78, p < .01).

Table 5: Results of PPs referring to humans using an R-pronoun. RP = R-pronoun and preposition attached, R … P = R-pronoun and preposition separated.

	Relative clause		**Topicalised PP**	
	number	**percentage**	**number**	**percentage**
RP	22	52,4	8	8,7
R … P	20	47,6	84	91,3
Total	**42**	**100**	**92**	**100**

The data are further analyzed in section 4.2.

4.2 Analysis

The data were analyzed using a logistic regression analysis. For these analyses some more items were excluded. All sentences belonging to compartment o in the CGN (see tables 1 and 2) were removed. Compartment o consists of read speech, it can therefore not be considered real spoken language. These were 7 sentences for the relative clauses and 9 sentences for the topicalised PPs. It is interesting to note that all but one of these had a regular PP (no R pronoun), and the one that didn't was clearly trying to reflect spoken language. Ten more sentences were excluded from the topicalised PPs because it was unclear what NP the pronoun referred to and/or its reference was not mentioned in the conversation.

The dependent variable was whether or not an R-pronoun was used: Its value is 0 if the sentence contains a regular PP and 1 if an R pronoun is used.

These were the independent variables I used:
Type: Relative clause or topicalised PP;
Region: The Netherlands or Flanders;
Distance: The distance between the (R-)pronoun and the NP it refers to in number of utterances (zero if it is in the same utterance);
Number: singular or plural;
Definiteness: definite or indefinite;

Complex: This variable is 1 if the NP consists of more than simply a determiner, adjective and noun. This includes for example NPs with a relative clause, but also NPs with coordination (using 'and' or 'or');
Pronoun: Whether the reference is pronominal. These are mostly words like 'all', 'someone', 'nothing'. There were no personal pronouns in the data, as explained above;
Name: Whether the NP is a name;
Celebrity: Whether the NP refers to a celebrity. This includes some names, but also words like 'singer' or 'composer'. In other words, they do not have to refer to one specific celebrity;
Group: For NPs that refer to (a group of) humans in a less direct way. These are words like 'group' or 'family'.

I used a forward stepwise procedure to find the best fitted model for the data. It includes the variables *Definiteness, Group* and *Distance* (all $p < .05$). This model predicts 69.9 % of the data correctly, while a baseline model, which simply predicts the most common outcome for every item, scores 67.7 % correct. For a summary of the model, see Table 6. According to these results, a pronoun is most likely to become an R-pronoun if its reference is indefinite, a group and not too many utterances away. Especially the effect of the group variable is strong: The odds of the pronoun being an R pronoun are over five times as high when the referent is a group than when it is not. When the referent is definite, the pronoun is about half as likely to become an R pronoun than it is with an indefinite referent. These two variables were somewhat correlated (X^2 (1, n = 186) = 5.4, p = .02), it is possible this influences the effects. The effect of *Distance* is almost one. The effects of other variables were not significant.

Table 6: Logistic regression analysis. Dependent variable: Construction (regular PP or R pronoun). All variables $p < .05$. Model: χ^2 (3) = 24,92, $p < .01$. R^2 = .13 (Cox & Snell), .18 (Nagelkerke).

		95 % CI for Odds Ratio		
	B (SE)	Lower	Odds Ratio	Upper
Constant	1,390 (0,344)			
Definiteness	-0,817 (0,384)	0,208	0,442	0,939
Group	1,633 (0,763)	1,148	5,118	22,825
Distance	-0,083 (0,033)	0,863	0,92	0,981

A second logistic regression analysis was carried out on the part of the data that contain an R pronoun. These included 42 sentences with relative clauses and 84 of the topicalised PPs. The goal of this test was to find out which factors are significant in whether or not the preposition moves along with the R-pronoun. The dependent variable here is therefore whether or not the R-pronoun and the preposition are attached (1 if they are, 0 if they are not). The best fitted model (again using the forward stepwise procedure) includes *Type* ($p < .01$) and *Celebrity* ($p < .05$). It correctly predicts 77.8 % of the data, compared to 76.2 % for the baseline model. In the topicalised sentences, the R pronoun and preposition are almost 15 times as likely to be separated than in the relative clauses. Also, when the pronoun refers to a celebrity, it is nearly five times as likely to be attached than when it does not. For a summary of the results, see Table 7.

Table 7: Logistic regression analysis. Dependent variable: Separation (R and P attached or not). Type: $p < .01$, Celebrity: $p < .05$. Model: χ^2 (2) = 32,40, $p < .01$. R^2 = .23 (Cox & Snell), .34 (Nagelkerke).

		95 % CI for Odds Ratio		
	B (SE)	Lower	Odds Ratio	Upper
Constant	0,001 (0,315)			
Type	-2,681 (0,550)	0,023	0,068	0,201
Celebrity	1,585 (0,703)	1,23	4,881	19,363

Interaction effects were not very reliable due to the low amount of data, so they were disregarded altogether.

5. Discussion

The aim of this thesis is to find out which factors influence the decision to use an R-form to refer to a human referent, focusing on the properties of the referent. I have investigated this using relative clauses, because this phenomenon is relatively well-known for relative clauses. In the Introduction, I suggested a reason for why it happens with this type of sentence specifically. The motivation for moving the pronoun in this case is wh-fronting, which is obligatory in Dutch. This causes the preposition to strand and the pronoun to become an R-pronoun (25a). Leaving the pronoun with its preposition is not an option, and stranding the preposition without turning the pronoun into an R-pronoun is ungrammatical as well. The only way to avoid referring to a human with an R-pronoun is to front the preposition along with the relative pronoun (pied-piping), as in (25b).

(25) a. Trio .
Trio
waar Eelko een bedrijfje **mee** heeft .
where Eelko a company **with** has
"Trio, whom Eelko owns a company with."
CGN: fn000434.208-209

b. d'r zijn wel mensen **bij** **wie** 't mooi staat .
there are PART people **at** **whom** it pretty stands
"There are people on which it looks good."
CGN: fn000667.87

The other construction that was used in this paper is a main clause in which the pronoun of the PP is fronted (26a). The reason for including this construction was Bouma's (2008) finding that a whole PP (as in (26b)) is relatively uncommon in a sentence-initial position. In addition to that, however, it parallels the relative clause in that the fronting of the pronoun here is very much obligatory. In this case, of course, leaving it in its unmarked position is grammatical, but it changes the meaning of the sentence significantly. Here, too, the only way to avoid an R-pronoun is by pied-piping the preposition (26b).

(26) a. en z'n vriendin was een blonde **daar** heb 'k ook
and his girlfriend was a blonde **there** have I also
nog koffie **mee** gedronken nog .
PART coffee **with** drunk PART
"…and his girlfriend was a blonde, I had a coffee with her too."
CGN: fn000948.165

b. dus **voor** **hem** was dat een keerpunt in zijn leven
so **for** **him** was that a turning-point in his life
"So for him, that was a turning point in his life."
CGN: fn009128.78

The reason for moving the pronoun in these two constructions is stronger than the scrambling in an unmarked main clause, therefore it is capable of (partly) overruling the constraint against combining R-forms with humans.

The results of the corpus study show that using an R-form to refer to a human is indeed quite common for both of these sentence types: For both types, the version with the R-form occured more often than the regular PP, although this difference was significant only for the topicalized sentences. This suggests topicalisation is better at overruling the humanness constraint than wh-fronting. In the case of relative clauses, however, adherence to the prescriptive norm could work in favour of the variant without an R-pronoun. As mentioned in section 3, the issue of referring to humans with R-pronouns receives much more attention with relative clauses than with other types of sentences. The difference between the two types did not come out as significant in the logistic regression analysis, however. When forced into the model it had a p of around .06; possibly future research can show whether there actually is a difference or not.
A closer look at the R-pronoun data only shows a more clear effect. Here, again, the difference is significant only for the topicalised sentences: the pronoun and preposition are a lot more likely to be disconnected. In order to explain this, we need to look at what the difference really is between these two types:

(27) a. **daarachter** zit Franck Vanderbroecke
there-behind sits Franck Vanderbroecke
"Behind him is Franck Vanderbroecke."

CGN: fn007162.3

b. en z'n vriendin was een blonde **daar** heb 'k ook
and his girlfriend was a blonde **there** have I also
nog koffie **mee** gedronken nog
PART coffee **with** drunk PART
"…and his girlfriend was a blonde, I had a coffee with her too."

CGN: fn000948.165

In sentence (27a), the preposition has pied-piped along with the pronoun to a sentence-initial position, but it still follows the pronoun which is therefore still an R-pronoun. In (27b), the preposition has remained in situ. I have argued that a resistance against pied-piping is the reason R-forms are used more often for humans in these sentences. If that is true, it does not make sense to pied-pipe the preposition while still keeping the R-pronoun. This seems to be true for the topicalisation data where (27a) consists of less than 10% of the R-pronoun data, but for the relative clauses both types occur equally. Apparently for these clauses, it is not only the pied-piping itself that matters, but there is some extra step between the two ways of pied-piping the preposition: That of moving it to before the pronoun (unlike (27a)) and in that way allowing it to remain a regular pronoun.
The exact reason why using an R-form for humans in these two constructions is not the focus of this thesis, however. I have shown that the variation between using an R-pronoun and a regular pronoun definitely exists in these data. The goal of the logistic regression analysis was to find out whether the properties of the referent can predict this variation. It is possible that for some types of human referents, the constraint on not using R-forms with humans is stronger than for others, and that therefore for some types, the preposition is moved to before the pronoun in order to stop it from becoming an R-pronoun. This is indeed the case. One variable that turned out to be significant is the one labelled *Group*. I used this variable for referents that refer to a group of humans, mostly words like 'group' and 'family':

(28) a. en dan K's Choice da 's *de* *enige* *groep* **waar** ik alle
and then K's Choice that's the only group where I all
CD's **van** koop .
cd's from buy
"and then there's K's Choice, that's *the only band* of which I buy every cd."

CGN: fv400195.96

b. *familie* **daar** moet je toch iets **voor** over hebben .
family there have you PART something for left have
"*Family* is something you should be willing to sacrifice things for."

CGN: fn000363.204

It is more acceptable to use an R-pronoun to refer to these group words than it is for other words. These words refer to a group as a whole. Although these groups consist of humans, it is not the individual that matters, it is the collective. The people in the group are just nameless faces, so to speak. This relates to Hopper & Thompson's (1980) notion of *Individuation*: "the degree to which the entity referred to by the NP is discrete, bounded, and separated from its environment" (Hopper & Thompson, 1980, p. 286). An individuated noun refers to a clearly defined entity, specific, preferably singular. A collective noun is the opposite of this. Hopper & Thompson found an effect of individuation on object marking: an individuated noun was more likely to receive explicit marking. I will return to what that means for these data below.

Another property that turned out to be significant is *Definiteness.* This is not very surprising: Definiteness has been shown to have an effect on many syntactic structures. Definite nouns are less likely to be associated with an R-pronoun than indefinite nouns. It would be easy to conclude that a definite noun is more 'human' than an indefinite noun. After all, animate nouns are more often definite than not - in the data for this thesis, too, almost two thirds is definite. De Swart (2007) however argues animacy and definiteness should be kept strictly separate. Animacy (therefore also humanness) is a lexical feature of a noun, it cannot be changed by adding a suffix, for example. Definiteness, on the other hand, is purely grammatical - no noun is inherently definite or indefinite. It does not make sense to say a definite noun is more human than an indefinite noun.
Why then does definiteness matter for an alternation based on animacy if definiteness and animacy are completely different things? Interestingly, definiteness too is a part of the idea of individuation (Hopper & Thompson, 1980). A definite noun is more individuated: It refers to a specific object (or group of objects), different from other objects. The reference of an indefinite noun is less specific, it is less of an individual.

Apparently, then, the more specific and individual the referent is, the less likely it is to be referred to with an R-pronoun. For these nouns, the humanness constraint against combining humans with R-forms is strong enough to overrule the unwillingness of the preposition to pied-pipe. For collective, less specific referents, the humanness constraint is violated more easily. This can be explained by politeness. An R-pronoun is a form normally meant for inanimate objects; therefore, using it for a human being is considered impolite, just as it is impolite to refer to a human as 'it'. Doing so is disrespecting of their humanness. The more specific and individuated a referent is, the more using an R-form will seem like a personal attack. When the referent is a collective, however, where the individuals are not important, no one will have to feel directly offended when referred to with an

inanimate reference word. In these cases there is therefore less incentive to avoid it, and the easy way out is taken by leaving the preposition where it is.

The third variable that had an effect was *Distance*. This property seems to be somewhat different from the other two. It is not so much a property of the noun phrase itself, as it is a property of its relation to the pronoun that refers to it. The further away the NP is from the pronoun, the more likely it is to be referred to using a personal pronoun. This could be a matter of discourse prominence (the further away the referent is, the less accessible it has become, and the speaker might want to emphasize its humanness again). Further research is required to confirm this effect.

As for the R-pronoun data only, the most significant effect was that of sentence type. This is not surprising, considering the difference between attached and unattached prepositions was highly significant for the topicalised clauses and not at all for the relative clauses. A possible explanation for this effect has been discussed above.
Another variable that was significant for this data is the variable *Celebrity*. The preposition is more likely to be fronted along with the pronoun (i.e. attached) when its referent is a celebrity. *Celebrity*, too, relates to *Individuation* (Hopper & Thompson, 1980): A reference to a celebrity is less direct, it is not usually someone the speaker knows personally. Also, often, by referring to a celebrity, the speaker is actually talking about something that celebrity is known for (e.g., their music). It could therefore be expected to have an effect on the likelihood of a pronoun becoming an R-pronoun. This, however, was not the case. The effect that I have found here, is on whether or not the preposition pied-pipes along with the pronoun, given that it is an R-pronoun. This effect was unexpected. Within this thesis, I do not have an explanation for it. Future research is needed to find the cause for this effect.

Finally, it should be noted that the two sentence types discussed in this thesis are not the only ones in which referring to humans with an R-pronoun occurs. They were chosen for this thesis because there is reason to assume it happens more often with these sentences, and therefore any possible effect is easier to find. I can however not even be certain that these are the constructions in which it occurs the most, because I have not compared them to any other type of sentence. Now that apparently there is an effect of the referent indeed, other types of sentences should be investigated to see whether the effect holds.

6. Conclusion

The aim of this thesis was to find out whether the properties of a human referent affect its chances of being referred to with an R-pronoun, despite the fact that R-forms generally resist being used for animates. This was done using the Spoken Dutch Corpus (CGN). Two types of sentences were included which both involved preposition stranding with an R-pronoun: relative clauses and main clauses with a topicalized PP. These were chosen because in these types both variants (with and without an R-pronoun) occur.
Referring to humans with R-pronouns turned out to be quite common: In the relative clauses, both forms were equally common, and in the topicalised clauses, the variant with an R-pronoun occurred significantly more often than the variant without an R-pronoun. The properties of the referent are shown to be able to predict the probability of it being referred to with an R-form: Collective and indefinite nouns are most likely to have the humanness constraint violated. These referents are less specific, therefore there is less need to be polite by respecting their humanness. The constraint on not using R-forms with humans can then be violated in favour of a constraint working in the opposite direction: Leaving the preposition where it originated (thus stranding it).

7. References

Aristar, A. (1996). The relationship between dative and locative: Kuryłowicz's argument from a typological perspective. *Diachronica*, 13(2), 207-224.

Van Bergen, G. & De Swart, P. (2010). Scrambling in spoken Dutch: Definiteness versus weight as determinants of word order variation. *Corpus Linguistics and Linguistic Theory*, 6(2), 267-295.

Van Bergen, G. (2011). *Who's first and what's next. Animacy and word order variation in Dutch language production.* PhD dissertation. Nijmegen: Radboud University Nijmegen.

Bouma, G. (2008). *Starting a sentence in Dutch: A corpus study of subject- and object-fronting.* PhD dissertation. Groningen: Rijksuniversiteit Groningen.

Broekhuis, H. (2002). Adpositions and adpositional phrases. *MGD Ocasional papers*, 3.

CGN (2006). *Corpus Gesproken Nederlands*, version 2.0. Electronic resource. See http://www.inl.nl/tst-centrale/images/stories/producten/documentatie/cgn_website/doc_English/start.htm.

Comrie, B. (1989). *Language universals and linguistic typology: Syntax and morphology.* Oxford: Blackwell.

Van Eynde, F. (2004). *Part of Speech-tagging en lemmatisering van het Corpus Gesproken Nederlands.* Leuven. Part of the documentation with CGN (2006).

Haeseryn, W., Romijn, K., Geerts, G., De Rooij, J. & Van den Toorn, M. (1997). *Algemene Nederlandse Spraakkunst*. Groningen: Nijhoff and Deurne: Wolters Plantyn.

Helmantel, M. (2002). *Interactions in the Dutch adpositional domain.* PhD dissertation. Leiden: Universiteit Leiden.

Hoekstra, H., Moortgat, M., Renmans, B., Schouppe, M., Schuurman, I. & Van der Wouden, T. (2003). *CGN syntactische annotatie.* Part of the documentation with CGN (2006).

Hoffman, T. (2005). Variable vs. categorical effects: Preposition pied piping and stranding in British English relative clauses. *Journal of English Linguistics*, 33(3), 257-297.

Hopper, P.J. & Thompson, S.A. (1980). Transitivity in grammar and discourse. *Language*, 56(2), 251-299.

Johansson, C. & Geisler, C. (1998). Pied-piping in spoken English. In A. Renouf (Ed.), *Explorations in corpus linguistics* (pp. 67-82). Amsterdam: Rodopi.

Kuno, S. (1976). Subject, theme, and the speaker's empathy: A re-examination of relativization phenomena. In C.N. Li (Ed.), *Subject and topic* (pp. 417-444). London/New York: Academic Press.

Kuryłowicz, J. (1964). *The inflectional categories of Indo-European.* Heidelberg: Carl Winter Universitätsverlag.

Landis, J.R. & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33 (1), 159-174.

Lestrade, S., De Schepper, K., Westelaken, S. & De Hoop, H. (2010). Preposition stranding everywhere. *ReVEL*, 8 (4), 223-247.

Van Riemsdijk, H. (1978). *A case study in syntactic markedness: The binding nature of prepositional phrases.* PhD dissertation. Amsterdam: University of Amsterdam.

Schippers, R. (2012). *More on preposition stranding without r-pronouns.* Presentation held at the TIN-dag, Utrecht, February 4th 2012.

De Swart, P. (2007). *Cross-linguistic variation in object marking.* PhD dissertation. Nijmegen: Radboud University Nijmegen. LOT Publications.

De Vries, M. (2006). Possessive relatives and (heavy) pied-piping. *Journal of Comparative Germanic Linguistics*, 9, 1-52.

Yamamoto, M. (1999). *Animacy and reference: A cognitive approach to corpus linguistics*. Amsterdam/Philadelphia: John Benjamins.

The lyrics in (1) are from the song 'Genoeg', by Eefje de Visser, the first song off her album *De koek* (2011).

Appendix A: Tiger syntax of the searches used in the CGN

(15a) het meisje **waarmee** ik danste
the girl **where.with** I danced
"The girl I danced with"

```
#n1: [cat="REL"] &
#n2: [pos="BW"] &
#n1 >RHD #n2
```

(15b) het meisje **waar** ik **mee** danste
the girl **where** I **with** danced
"The girl I danced with"

```
#n1: [cat="REL"] &
#n2: [pos="VZ2"] &
#n1 >* #n2 &
#n3: [cat="PP"] &
#n3 >HD #n2 &
#n1 >RHD #n4:[pos="VNW15"]
```

(15c) het meisje **met** **wie** ik danste
the girl **with** **who** I danced
"The girl with whom I danced"

```
#n1: [cat="REL"] &
#n1 >RHD #n4:[cat="PP"]
```

(16a) **daarmee** heb ik altijd ruzie
there.with have I always fight
"I'm always fighting with him/her!"

```
#n1: [cat="SMAIN"] &
#n1 >HD #n2 &
#n1 >* #n3 &
#n3: [word="daarmee"] &
#n3 .* #n2
```

(16b) **daar** heb ik altijd ruzie **mee**
there have I always fight **with**
"I'm always fighting with him/her!"

```
#n1: [cat="SMAIN"] &
#n1 >HD #n2 &
#n1 >* #n4:[cat="PP"] &
#n4 >OBJ1 #n3 &
#n3:[pos="VNW20"] &
#n3 .* #n2
```

(16c) **met** **hem** heb ik altijd ruzie
with him have I always fight
“I’m always fighting with him!”

```
#n1: [cat="SMAIN"] &
#n1 >HD #n2 &
#n1 >* #n4:[cat="PP"] &
#n4 >OBJ1 #n3 &
#n3:[pos="VNW2"] &
#n4 >HD #n5 &
#n5 .* #n2
```

Appendix B: List of prepositions used to search for topicalised R-pronouns with the preposition attached

1. mee "with"
2. aan "at, on"
3. voor "for, before, in front of"
4. op "on, on top of"
5. van "from, of"
6. uit "out (of)"
7. achter "behind"
8. bij "near, at"
9. over "over"
10. naartoe "to"
11. heen "to"
12. in "in"
13. naast "next to"
14. buiten "except"
15. langs "alongside"
16. na "after"
17. omtrent "about, concerning"
18. tegen "against"
19. toe "to"
20. tussen "between"
21. af "off"
22. binnen "within"

Appendix C: Some sentences with annotation

Type		Number	Session	Utterance	Reference									Preposition	Language	Compartm
						Nr of utte	Distance	Number	Definiten	Pronoun	Complex	Name	Celebrity	Less human		
PP	0	4	fn000667.	d'r d'r zijn	mensen		0	1	0	0	0	0	0	0 bij	N	a
PP	0	20	fv400112.	maar er zi	de leerlin	114	2	1	1	0	1	0	0	0 bij	V	b
R ... P	1	50	fn000389.	en toen li	zo'n jongen die bij m		0	0	1	0	1	0	0	0 tegen	N	a
R ... P	1	6	fn000042.	uh met na	Uphoff		0	0	1	0	0	1	1	0 van	N	f
RP	1	11	fn000254.	[daarvoor	een assist	143	3	0	0	0	0	0	0	0 voor	N	a