

A comparison between *als* and *dan* in comparative constructions of inequality

Ferdy Hubers
s3011682
22nd of August, 2012
Bachelor thesis
Prof. Dr. Helen de Hoop

Table of contents

TABLE OF CONTENTS	II
1 INTRODUCTION	1
2 FACTORS STIMULATING THE USE OF <i>ALS</i> AS A COMPARATIVE MARKER	2
2.1 INTRODUCTION	2
2.2 THE CONTEXTS AND GRAMMATICAL FUNCTIONS OF <i>ALS</i> AND <i>DAN</i>	2
2.3 PREPOSITIONS IN COMPARATIVE CONSTRUCTIONS	4
2.4 PROPERTIES OF CONJUNCTIONS AND PREPOSITIONS	6
2.5 DISCUSSION	7
3 <i>ALS</i> AND <i>DAN</i> AS COMPARATIVE MARKERS: A CORPUS STUDY	9
3.1 INTRODUCTION	9
3.2 METHOD	9
3.2.1 Preparation	9
3.2.2 Annotation of the data	11
3.2.3 Random sample of the data	12
3.2.4 Variables for statistical analysis	13
3.2.5 Statistical analysis	14
3.3 RESULTS	15
3.3.1 Crosstabs	15
3.3.2 Logistic regression	18
3.4 CONCLUSION	19
4 DISCUSSION	21
5 CONCLUSION	22
6 REFERENCES	23

1 Introduction

Over decades people have been arguing about the use of *als* 'as'¹ and *dan* 'than' in comparative constructions of inequality in Dutch. According to Stroop (2010) *dan* 'than' as a comparative marker was very common in Middle Dutch and the construction *groter als* 'bigger as' was not known yet. Around the 16th century the use of *als* 'as' as a comparative marker arose and became very common in a short period of time. It even occurred in the written texts of prominent writers like van den Vondel and Huygens (Stroop, 2010). In their oeuvre they used as much *als* 'as' as *dan* 'than' in comparative constructions. This use of *als* 'as' came from the dialects of Dutch in which it was quite normal to use *als* 'as' in comparative constructions. In 1720 the linguist Balthazar Huydecoper thought of the comparative marker *als* 'as' as corruption of the language. He argued that everyone had to speak and write like in the Middle Ages, because Middle Dutch was not yet infected by loans and therefore was better (Stroop, 2010). In Middle Dutch *als* 'as' as a comparative marker of inequality was not yet in use, so it had to be banned from Dutch. Because of the status of Huydecoper his followers also tried to decrease the use of *als* 'as' in comparative constructions. This resulted in a prescriptive grammar in which *dan* 'than' was prescribed as the comparative marker in comparative constructions of inequality. In the 20th century the linguist Paardekooper (1950, 1970) argued that the use of *als* 'as' as a comparative marker was originally Dutch, because it came from the old Dutch dialects. Therefore he consequently used *als* 'as' instead of *dan* 'than' in comparative constructions in his own work.

Nowadays the prescriptive grammar still disapproves of *als* 'as' as a comparative marker. *Dan* 'than' is used as a comparative marker instead, like in *Jan loopt harder dan jij* 'Jan walks faster than you'. Still, in spontaneous conversations such an utterance is produced many times with *als* 'as' instead of *dan* 'than'. Why do some people use *als* 'as' as a comparative marker in comparative constructions of inequality even though the prescriptive grammar forbids it? In this Bachelor's thesis I will try to determine what reasons speakers have to use *als* 'as' and *dan* 'than' as a comparative marker in comparative constructions of inequality.

On the basis of grammar-internal factors *als* 'as' might be more likely to occur in a comparative construction than *dan* 'than'. These grammar-internal factors consist of the idea of analogy and economy. On the basis of the prescriptive rule learned at school *dan* 'than' is more likely to occur in comparative constructions. This is called the grammar-external factor. Perhaps other factors too stimulate the use of *dan* 'than'.

In Section 2 I will discuss the grammar-internal factors that might be of influence in the prediction of *als* 'as' versus *dan* 'than' in comparative constructions. In the third section I will look into more detail to the grammar-external explanation by means of a corpus study. Here I will also present its results. In the fourth section I will interpret the results and this Bachelor's thesis ends with a general conclusion.

¹ Throughout this thesis, I will only give the English translation of *als* and *dan* in section 1, in order to keep the text readable.

2 Factors stimulating the use of *als* as a comparative marker

2.1 Introduction

Als and *dan* are used in all kinds of contexts in Dutch. An example of a context we have already seen is the context of comparative constructions. *Als* also occurs in conditionals for example, whereas *dan* occurs in contexts in which it is used as a sentence adverb. In these contexts *als* and *dan* both have a certain grammatical function. For *als* it is possible to occur as a conjunction. As I already mentioned *dan* can be used as a sentence adverb. In comparative constructions *als* and *dan* both have the same function. By looking at the grammatical functions of *als* and *dan* in other contexts it will become clear why this function is important in explaining the occurrence of *als* in comparative constructions in which *dan* is prescribed. In subsection 2.2 the contexts in which *als* and *dan* occur will be discussed as well as their grammatical functions in the different contexts. In subsection 2.3 I will argue that a comparative marker, besides a conjunction according to the traditional view, also can be regarded as a preposition. Section 2.4 will show how hard it is to distinguish conjunctions and prepositions. I will end Section 2 with a discussion.

2.2 The contexts and grammatical functions of *als* and *dan*

Example (1) shows some Dutch utterances in which *als* occurs. These utterances are taken from the components a and b of the Spoken Dutch Corpus (Oostdijk, 2003). Component a contains utterances from spontaneous face-to-face conversations and component b includes spontaneous speech from interviews with teachers of Dutch. It is possible for *als* to occur as a conjunction introducing a conditional clause, as can be seen in examples (1a) and (1b). Besides this use of *als* it can also occur in sentences like in (1c) and (1d). Here *als* stands for *in the capacity of* or *in the position of*. Here *als* can be regarded as a preposition.

(1)

- a. Maar **als** je 't vak echt op de rails wil zetten dan moet je d'r ook wat van weten. [comp. B: fn000109.69]²
'But **if** you want to get the course on the rails then you have to know something about it'.
- b. **Als** er niet snel wordt ingegrepen zal de gemeente het budget van 1,5 miljoen gulden teruggeisen. [comp. A: fn000003.11]
'**If** not intervened quickly the municipality will reclaim the budget of 1,5 million guilders'.
- c. Want ik stel me voor je bent er **als** boer uit of niet uit.
[comp. B: fn000096.288]
'Because I imagine, **as** a farmer you know it or you don't'.
- d. Ook **als** docent zijnde van het VBO krijg je nog een etiketje opgeplakt op je voorhoofd. [comp. B: fn000108.140]
'Also **as** a teacher of the VBO you get labelled'.
- e. Nou dat is vast net zo erg **als** The Bold And The Beautiful.
[comp. B: fn000005.319]
'Now, that surely is just as bad **as** The Bold And The Beautiful'.
- f. iets kleiner **als** een halve liter maar iets groter **als** een glas bier
[comp. A: fn000325.58]
'a bit smaller **than** a half a litre but a bit bigger **than** a glass of beer'

² Throughout this thesis, the examples are taken from the CGN (Spoken Dutch Corpus). For the sake of clarity I left out those hesitations and particles (the latter being used a lot in Dutch) that might blur the intention of the speaker. In all cases, the CGN code is given, which makes it possible for the reader to recover the original utterance in the CGN.

- g. maar zijn wel een beetje zachter **als** dat ze moeten zijn
 [comp. A: fn000293.121]
 'but are a little softer **than** they have to be'

Als is also used in comparisons, like in the examples (1e-g). In (1e) two entities are compared which are equal and traditionally *als* is regarded as a conjunction. The sentences in (1f) and (1g) compare two entities that are not equal. In these cases *als* is considered to be a conjunction too. In the next subsection I will argue that *als* in (1e) and (1f) is a preposition, however, and not a conjunction.

According to the prescriptive grammar of Dutch *als* used as a comparative marker in comparative constructions of inequality is not allowed. In this situation the speaker or writer should use *dan*. The *taaladviesdienst* of the *Nederlandse Taalunie* states:

'Volgens de traditionele schoolregels komt er na een vergrotende trap *dan* (*groter dan, liever dan*). *Dan* duidt erop dat we bij een vergelijking een ongelijkheid vaststellen.' (de Nederlandse Taalunie, Taaluniversum)

'According to the traditional rules learned at school *dan* is used after a comparative (bigger than, nicer than). *Dan* indicates that in a comparison an inequality is encountered.'

In example (2) the sentence from (1f) can be seen with *dan*. The meaning of the utterance does not change, but according to prescriptive grammar this is a correct utterance in contrast to the utterance in (1f).

- (2) iets kleiner **dan** een halve liter maar iets groter **dan** een glas bier
 'a bit smaller **than** a half a litre but a bit bigger **than** a glass of beer'

The use of *als* in constructions like (1f) and (1g) is very frequent in Dutch. In the Spoken Dutch Corpus this use of *als* is found 227 times out of the 1465 comparisons of inequality. This means that in 15.5% *als* is used in these constructions instead of the prescribed *dan*.

Since I have discussed all types of occurrences and grammatical functions of *als*, I will now take a look at the occurrences of *dan* in the components of the Spoken Dutch Corpus which contain spontaneous speech. In example (3) these occurrences are shown.

- (3)
- a. Nou als dat de helft zou zijn **dan** zou dat een stuk prettiger zijn.
 [comp. B: fn000004.42]
 'Now, if that would be half of it, **then** it would be much nicer.'
 - b. **Dan** doe je 't te goed. [comp. A: fn000508.134]
 'Then you do it too well.'
 - c. Ik kan **dan** bij m'n ouders slapen. [comp. A: fn000838.256]
 'I can sleep at my parents' home **then**.'
 - d. Dus soms dan zie je een beek dwars door 't bos lopen en **dan** komt ie uit een bos en **dan** worden 't weilanden aan weerskanten of aan één kant.
 [comp. A: fn000521.69]
 'Thus, sometimes you see a brook flow through the forest and **then** it comes out of the forest and **then** meadows appear on both sides or on one side.'
 - e. 't is echt zo'n soort Fredbus maar **dan** is ie vuurrood.
 [comp. A: fn000828.72]
 'It really is a kind of Fredbus, but **then** it is scarlet.'
 - f. Dan heeft ie twee banken een tafeltje en **dan** nog een eettafel.
 [comp. A: fn000626.193]
 'Then he has two couches, a table and **then** also a dining table.'
 - g. Hoe ver **dan**? [comp. A: fn000983.123]
 'How far **then**'

- h. En tegen de tijd dat hij veertig is verdient ie veel meer **dan** deze vent.
[comp. A: fn000509.80]
'And by the time he turns forty he earns much more **than** this guy.'
- i. Nou d'r zijn d'r wel net iets meer gestopt **dan** ik eigenlijk had gehoopt.
[comp. C: fn008314.152]
'There are more people quitted **than** I had actually hoped.'

It is possible for *dan* to occur as a sentence adverb, as can be seen in the examples (3a) and (3b). In (3c) and (3d) *dan* is used as a temporal adverb. In (3c) it has a punctual meaning (van Bergen, 2010). In this case *dan* is used to point to a time in the future. In (3d) *dan* is used to express a sequence of events (van Bergen, 2010). The sentence in (3e) is an example of *dan* used as a particle, because its meaning is not quite clear. The same goes for *dan* in the sentences (3f) and (3g). Here it does not really have a grammatical function either. In the last two examples, (3h) and (3i), *dan* is used in comparative constructions in which two entities are compared that are not equal. Traditionally, just like in comparative constructions with *als*, *dan* is seen as a conjunction only in this context. However, in cases like (3h) it actually is a preposition. The same goes for the cases like (1e) and (1f).

2.3 Prepositions in comparative constructions

Example (4) shows two types of sentences with a comparative construction. The sentences look very similar, but they differ in the complement of *als*. I will use the terms *clausal* and *phrasal comparative* to refer to the different types of comparative constructions (Hendriks, 1995). An example of a *clausal comparative* can be found in (4a). This is a *clausal comparative* because the complement of *als* consists of a clause. In (4b) a *phrasal comparative* is shown. In the case of a *phrasal comparative* the complement of *dan* or *als* consists of a single phrase. In example (4b) this single phrase is the determiner phrase (DP) *haar zus* 'her sister'.

- (4)
 - a. Zij is groter dan haar zus is.
'She is taller than her sister is.'
 - b. Zij is groter dan haar zus.
'She is taller than her sister.'

According to Hendriks (1995) the more traditional view on comparative markers views comparative constructions like (4b) as elliptic sentences, which means that they are reduced variants of clausal comparatives. The underlying form of (4b) thus is the sentence in (4a). This traditional view on comparative constructions serves as evidence for the fact that only conjunctions occur in comparative constructions, because when the underlying form is a clausal comparative the complement of the comparative marker is an IP. When the complement of *als* or *dan* is an IP (a clause) they can be regarded as conjunctions and not as prepositions since their complement is not a single phrase. However, Pinkham (1985) argues that in general phrasal comparatives are derived from clausal comparatives, but that there are a few exceptions. One of the exceptions can be seen in example (5).

- (5)
 - a. Jan rent harder dan het wereldrecord.
'Jan runs faster than the world record.'
 - b. *Jan rent harder dan het wereldrecord rent.
*'Jan runs faster than the world record runs.'

In (5a) the complement of *dan* is the determiner phrase *het wereldrecord* 'the world record'. If a phrasal comparative were derived from a clausal comparative, it should be possible to transform this construction into a clausal comparative by adding a verb, as

can be seen in (5b). Here the verb *rennen* 'to run' is added, so the complement of *als* becomes an IP. However, this is not possible, because a world record cannot run. This is why (5b) is infelicitous.

Another example, which is equal to (5), is the example in (6a). Here again, the underlying form cannot be a clausal comparative, because 23 years does not get older. The underlying clausal comparative, which can be seen in (6b), is ungrammatical.

- (6)
- a. Hij werd niet ouder dan 23 jaar.
'He did not get older than 23 years.'
 - b. * Hij werd niet ouder dan 23 jaar werd.
* 'He did not get older than 23 years got old.'

Another reason for arguing against the fact that a phrasal comparative is derived from a clausal comparative is reflexivity (Hendriks, 1995; Hoeksema, 1983). The sentence in (7a) is a phrasal comparative which is completely acceptable in Dutch. Its counterpart in (7b), however, which contains a clausal comparative, is ungrammatical. Its ungrammaticality lies in the fact that *zichzelf* 'himself' is a reflexive pronoun that cannot serve as a subject. The sentence in (7a) shows that it is well possible for a reflexive pronoun to occur as the complement of a comparative marker without assuming an underlying clausal comparative, since that clausal base turns out to be ungrammatical.

- (7)
- a. Niemand loopt harder dan zichzelf.
'No man runs faster than himself.'
 - b. * Niemand loopt harder dan zichzelf loopt.
* 'No man runs faster than himself runs.'

The final argument against the clausal basis of phrasal comparatives is stated by Hendriks (1995) and is about the malfunctioning of the rule of Comparative Ellipsis. This rule aims to delete some elements of the complement-clause of the comparative marker in a comparative construction, so the clausal comparative becomes a phrasal comparative (Pinkham, 1985). Hendriks argues that in some cases in Dutch deletion not only takes place at the complement of the comparative marker, but also at the phrase preceding this marker. An example from Hendriks (1995) is shown in (8).

- (8)
- a. Meer mensen hebben voor dan tegen het voorstel gestemd.
'More people have for than against the proposal voted.'
 - b. Meer mensen hebben voor het voorstel gestemd dan mensen tegen het voorstel gestemd hebben.
'More people have for the proposal voted than people against the proposal voted have.'

The sentence in (8a) is a phrasal comparative and should have been derived from (8b), which states the same meaning in a clausal comparative. The only way of getting from (8b) to (8a) is by deletion of elements preceding *dan* as well as by deletion of elements following *dan*. These deletion operations must be executed at the same time since the sentence in (9), in which only deletion of elements preceding *dan* has taken place, is ungrammatical. This means that the rule of Comparative Ellipsis does not work well in this case and that therefore this phrasal comparative does not have a clausal base.

- (9) * Meer mensen hebben voor dan mensen tegen het voorstel gestemd hebben.
* 'More people have for than people against the proposal voted have.'

The arguments above show that the distinction between phrasal and clausal comparatives is justified. When a comparative construction consists of a clausal comparative the comparative marker can be seen as a conjunction since in this case the complement of the comparative marker is a clause.

When a comparative construction is a phrasal comparative the comparative marker can be seen as a preposition since the complement of the comparative marker is a single phrase. A single phrase can be a determiner phrase (DP) or a prepositional phrase (PP). In example (4b) the complement of the comparative marker is a determiner phrase. Therefore the sentence can be seen as a phrasal comparative and thus the comparative marker is a preposition. In the following example the complement of the comparative marker is a prepositional phrase and thus can the sentence be regarded as a phrasal comparative. As I said, in this case the grammatical function of the comparative marker is a preposition too, because here we deal with a complex prepositional PP (Helmantel, 2002).

- (10) Hij kan beter aan tafel studeren dan op de bank.
'He better can study at the table than on the couch.'

The phrase *op de bank* 'on the couch' is a prepositional phrase which is the complement of the comparative marker *dan*. If *dan* here has to be seen as a preposition, two prepositions are juxtaposed. Consider an example from Helmantel (2002) in (11). This shows that it is very well possible for prepositions to be juxtaposed. In this example *van* 'from' is a preposition. The complement of *van* 'from' is the PP *achter het station* 'behind the station' and is of the same type as example (10). This is a complex prepositional PP too.

- (11) De bus vertrekt van achter het station.
'The bus departs from behind the station.'

2.4 Properties of conjunctions and prepositions

Prepositions originally have a spatial meaning. They connect the Ground and the Figure (Talmy, 2000). In the sentence in (12) *de pen* 'the pen' functions as Figure and *de kast* 'the closet' as Ground. The preposition *in* 'in' connects these two entities. The relation between Figure and Ground in this sentence is a spatial one. In the example in (13) not a spatial, but a temporal relation is present, while the same preposition is used (Haspelmath, 1997). *In* 'in' is used here to connect the clause *hij zwemt* 'he swims' and the phrase *de vakantie* 'the holiday'.

- (12) De pen ligt in de kast.
'The pen is in the closet.'
- (13) Hij zwemt in de vakantie.
'He swims during the holiday.'

In (14) the preposition *tot* 'to' has a spatial meaning. In (15) the same preposition *tot* 'until' has a temporal meaning since it indicates until what time the person works. Clausal conjunctions do not connect two entities or objects, but they only connect two events. By connecting two events the type of relation cannot be spatial. Events occur in a certain period of time. The relation expressed by the conjunction thus is temporal. An example can be found in (16). Here the conjunction *tot* 'until' connects the two events *hij fietste* 'he cycled' and *hij viel er bij neer* 'he fell'.

A conjunction is also very suitable for indicating a causal relation between two events. The reason for this is that people are very likely to interpret two successive events as a cause-effect as can be seen in (17). The conjunction *sinds* 'since' tells us that from the moment the boy fell, he was sick. This can very easily be interpreted as a cause-effect. *Hij was gevallen* 'he fell' then becomes the cause of *hij was ziek* 'he was sick'.

- (14) Hij fietste tot de brug.
'He cycled to the bridge.'
- (15) Hij werkt tot vijf uur.
'He works until five o'clock.'
- (16) Hij fietste tot hij er bij neerviel.
'He cycled until he fell.'

- (17) Hij was ziek sinds hij was gevallen.
'He was sick since he fell.'

From above it can be seen that prepositions and conjunctions semantically show some overlap. Prepositions can be used to express a temporal relation as can conjunctions (Haspelmath, 1997). Because of this similar function of expressing a temporal relation it is possible for a preposition to become a conjunction. This is what happened with the English word *since*. When looking at the syntax, it can be used as a preposition, as can be seen in example (18a). Here a temporal relation is expressed. It can also be a conjunction which expresses a temporal relation, as is shown in example (18b). It is even possible for *since* to function as a conjunction which expresses a causal relation. This is shown by example (18c). The same goes for *before* as can be seen in (19a-c).

- (18)
- a. Since the war, I am much more careful.
 - b. I lived here since he left.
 - c. I hope he will come back, since I feel lonely.
- (19)
- a. Before the house is a parking area.
 - b. Before the war, he fell ill.
 - c. Before he fell ill, he was a teacher.

The overlap in function between preposition and conjunction has a cross-linguistic nature. In French for instance this overlap is present too. Example (20a) shows *depuis* 'since' as a preposition which expresses a temporal relation. In (20b) the same word *depuis* 'since' is used as a conjunction which also expresses a temporal relation. In contrast to English, French does not have *depuis* 'since' used to indicate a causal relation.

- (20)
- a. Depuis la guerre, je suis plus prudent.
'Since the war, I am much more careful.'
 - b. Depuis la guerre est finie, je me sens sûr encore.
'Since the war is over, I feel safe again.'

The reason for the difficulty of distinguishing prepositions and conjunctions is the semantic overlap they show. Both prepositions and conjunctions are able to express a temporal relation. Therefore prepositions can often also be used as conjunctions. Syntactically, it is easy to distinguish prepositions and conjunctions. Within the category of preposition the objects or events they take as their complement are expressed by DPs as can be seen in the examples (12) and (13). *De pen* 'the pen' and *de kast* 'the closet' both are DPs as well as *de vakantie* 'the holiday'. In case of conjunctions the events are expressed by clauses, as the examples (16) and (17) show. The complements *hij erbij neerviel* 'he fell' from example (16) and *hij was gevallen* 'he fell' from example (17) are both clauses (IPs).

Due to the semantic overlap the difference between prepositions and conjunctions is not so clear. The only way of distinguishing them is to consider the syntactic properties of prepositions and conjunctions. The complement of a conjunction is a clause, whereas the complement of a preposition is a DP, or a PP, as I showed in the previous section.

2.5 Discussion

In Section 2 it has become clear that the grammatical category of the comparative marker in comparative constructions is a conjunction or a preposition. The comparative marker is a conjunction when dealing with a clausal comparative, so the complement of the comparative marker is a clause. The comparative marker can be regarded as a preposition when dealing with a phrasal comparative in which its complement is a single phrase (a DP or PP). Note that *als* has these two functions in other contexts too. However, *dan* is a preposition or a conjunction in the context of a comparative

construction only and corresponds in this case to the English word *than*. In other contexts it belongs to the classes of adverbs and particles and corresponds with the English word *then*.

Some grammar-internal factors stimulate the use of *als* as a comparative marker. Analogy and economy are at the basis of these factors. According to analogy *als* can easily be used as a comparative marker since *als* does not have any other grammatical functions besides conjunction and preposition. This is also more economic than using *dan* since this word has two more grammatical functions. *Als* is also already used as a comparative marker in comparative constructions of equality. This makes it suitable as a comparative marker in comparative constructions of inequality too. Besides that, the influence of dialects may stimulate the use of *als* as a comparative marker in comparative constructions, since many dialects use *als* in these comparative constructions.

The use of *dan* as a comparative marker seems a bit strange, because only in the context of comparative constructions *dan* is used as conjunction or preposition. In all other contexts it is as an adverb or particle. Despite the grammar-internal factors that nominate *als* as comparative marker, *dan* still is much more frequent in these contexts. This might be due to the prescriptive rule learned at school that prescribes *dan* as a comparative marker in comparative constructions of inequality. At school it is taught that chunks like *groter dan* 'bigger than' are correct, therefore education should be of importance in predicting *als* and *dan* as comparative markers of inequality. When people use *dan* in comparative constructions because they learned chunks like *groter dan* 'bigger than' whereas grammar-internal factors stimulate the use of *als* in this context, a difference in use is expected. When no elements occur between the comparative element and the comparative marker, people will use *dan* since this is analogous to the chunk *groter dan* 'bigger than' that stands at the basis of the prescriptive rule. When the amount of elements between the comparative element and the comparative marker increases it might happen that people use *als* instead of *dan* due to the grammar-internal factors, because the chunk that stands at the basis of the prescriptive rule is not very similar to the comparative construction they want to produce. An example can be seen in (21).

(21)

- a. Hij is groter dan zij.
'He is bigger than her'.
- b. In de toekomst wil hij harder kunnen lopen als zij.
'In the future he wants to be able to walk faster than her.'

We might hypothesize that in (21a) *dan* will be more frequent as a comparative marker than in (21b), because the chunk is completely similar to the chunk that stands at the basis of the prescriptive rule. In (21b) the words *kunnen* 'can' and *lopen* 'walk' are used in between the comparative element *harder* 'faster' and the comparative marker, and people might not be able to recognize the construction right away as a comparative construction of inequality, and therefore use the 'incorrect' form *als*, due to the grammar-internal factors. In order to test whether indeed various factors influence and can be used to predict the choice between *als* and *dan* in comparatives of inequality I have conducted a corpus study based on the Spoken Dutch Corpus. I will present this study in Section 3 below.

3 *Als* and *dan* as comparative markers: a corpus study

3.1 Introduction

To examine the distribution of the comparative markers *als* and *dan* I have searched for utterances containing these words in combination with a comparative in the Spoken Dutch Corpus, which consists of standard Dutch as spoken by adults in The Netherlands and Flanders. I used COREX to do so. This tool is especially developed to search through this corpus. It is possible to search for a specific orthographic annotation as well as a part-of-speech tag or a prosodic feature.

For this study the Spoken Dutch Corpus is chosen, because it is well searchable since all data is transcribed orthographically. Besides that, the corpus is also syntactically annotated for about one million words. Another reason to use this corpus is the amount of data available. The corpus consists of nearly 9 million words. About 3.3 million words were collected in Flanders and over 5.6 million words in The Netherlands. All words come from spoken Dutch. This is a big advantage because the use of *als* in comparative constructions barely occurs in texts. The corpus includes all kinds of speech, like face-to-face conversations, but also monologues and read speech. In the next subsection I will describe the methodology. In subsection 3.2 I will present the results of this study and I will end this section with a conclusion to summarize the main results.

3.2 Method

3.2.1 Preparation

First, I had to know how *als* and *dan* were tagged in the corpus, in order to get data in which they function as a comparative marker. Therefore I searched in COREX by means of Tiger Search, which looks only for syntactic annotated data. As a result of that search it became clear that *als* and *dan* were not always tagged in the same way in comparative constructions. In most cases they were both tagged as a conjunction, but *dan* also was tagged as an adverb in these contexts.

To search for *dan* and *als* in comparative constructions only I also had to specify the grammatical tag of the comparative. As a result of Tiger Search it became clear that four tags were possible for comparatives. A comparative could be tagged as *ADJ10*, which stands for adjectives ending in *-er* like *groter* 'bigger' and *goedkoper* 'cheaper' as can be seen in the sentence in (22a). Comparatives were also tagged as *ADJ7* as example (22b) shows us. This stands for adjectives ending in *-s*. Examples are *anders* 'different' and *nuttig* 'useful'. The other types of comparatives consist of words like *meer* 'more' and *minder* 'less'. These types of comparatives are called *VNW24* and *VNW26*. In comparatives of type *VNW26* words like *meer* 'more' and *minder* 'less' occur directly in front of the comparative marker as in (22c) is shown. Comparatives that are marked as *VNW24* always occur in combination with a noun, example (22d), or adjective/adverb as in example (22e).

(22)

- a. En is ze dan nog goedkoper als een gewone kapper?
[comp. A: fn000351.183]
'And she is still cheaper than an ordinary hairdresser?'
- b. Jij had wat anders als ik hè?
[comp. A: fn000792.1]
'You had something different than me, hadn't you?'
- c. Ik werk vier uur per week meer als jij.
[comp. A: fn000845.65]
'I work four hours a week more than you.'
- d. Hij had ook wel meer capaciteiten dan Pierre Van Hooijdonk.
[comp. A: fn007886.17]
'He had more skills than Pierre Van Hooijdonk.'

- e. Kunt u omschrijven waarom bezit beroven van iemand volgens u minder zwaar doorweegt dan uhm het raken aan kinderen?
 [comp. H: fv400048.36]
 'Can you describe why you think robbing ones property is less bad than touching children?'

To summarize, two tags exist for *als* and *dan*, which can be seen in Table 1 below. For the comparatives four tags exist, which are listed in Table 2.

Table 1 Possible tags of comparative markers

Comparative marker	Tag in Spoken Dutch Corpus
Als	VG2
Dan	VG2; BW

Table 2 List of possible tags of comparatives

Tag in Spoken Dutch Corpus for Comparatives	Meaning
ADJ10	Adjectives with suffix <i>-er</i> ; <i>groter</i> 'bigger', <i>kleiner</i> 'smaller', <i>leuker</i> 'more nice'
ADJ7	Adjectives with suffix <i>-s</i> ; <i>anders</i> 'else', <i>nuttigs</i> 'useful'
VNW24	<i>Meer</i> 'more' and <i>minder</i> 'less' used independently
VNW26	<i>Meer</i> 'more' and <i>minder</i> 'less' used in combination with a noun or adjective/adverb.

Next, the distance between the comparative and the comparative marker had to be determined. If not limiting this distance, the search would produce a lot of false hits, since much data would be found in which *als* and *dan* do not function as comparative markers anymore. Zero up to two words regularly occur in between the comparative and the comparative marker as can be seen in (23a-b). Even three words is possible as examples (23c-d) show.

(23)

- a. Hij is niet verder gekomen als tweeduizend.
 [comp. A: fn000841.265]
 'He did not get further than two thousand.'
- b. Zo heb ik iets minder kunnen leren als andere jaren.
 [comp. B: fv400178.126]
 'Like this I less have learned than other years.'
- c. Vroeger kregen de kinderen veel meer huiswerk naar huis dan tegenwoordig. [comp. H: fn009223.49]
 'Before children got much more home work at home than nowadays.'
- d. erger voor de docent dan voor de studenten
 [comp. A: fn000576.100]
 'worse for the teacher than for the students'

Taken all these considerations into account the queries to search for the comparative constructions were formulated. I have searched in *content search* for *als* as VG2 'conjunction' in combination with all four types of comparatives. I have restricted the amount of words that can occur between the comparative and the comparative marker to a maximum of 3. I have done the same for *dan* with the tag VG2 'conjunction' as well as for *dan* with the tag BW 'adverb'. I collected 4084 comparative constructions in which *dan* is used as a comparative marker and 481 sentences that go with *als* as a comparative marker. This data had to be annotated. In the next section I will discuss what the annotation process looked like.

3.2.2 Annotation of the data

I annotated the comparatives and their tag for each utterance. I also annotated in what way the comparative was used. One possible annotation was *free*. An example of a free comparative can be found in (24). This tag is used in the Spoken Dutch Corpus in the same way.

- (24) en wat anders is dan twee jaar geleden, zo heb ik begrepen
[comp. G: fn000154.135]
'and that is different than two years ago, I thought'

The comparative *anders* 'else' is not used in combination with a noun or adjective. In (25) an example is given in which the comparative is in pre-adjectival position and in (26) it is pre-nominal.

- (25) misschien minder gedetailleerd als die van collega Dielens
[comp. G: fv600685.27]
'perhaps less detailed than the one from colleague Dielens'
(26) Met de één heb je meer contact mee als de ander.
[comp. A: fn000777.180]
'You have more contact with one person than with another.'

The comparative as encountered in (27) is annotated as *free*. In this case maybe *preverbal* should have been a better name, but I thought of this comparative in the same way as the comparative in (28). In (28) the comparative *beter* 'better' is *free*. Analogous to (28), the comparative in (27) also has to be considered as *free*, because both sentences express exactly the same proposition.

- (27) nee die zich beter voelen als een ander omdat ze meer geld hebben
[comp. A: fn000616.130]
'no who feels better about himself than about another because they have more money'
(28) nee die zich beter als een ander voelen omdat ze meer geld hebben
'no who feels better about himself than about another because they have more money'

According to the hypothesis formulated at the end of Section 2 I expect the distance between the comparative and the comparative marker is of importance in using *als* over *dan* as a comparative marker. If an effect of distance would be present, this indicates that people apply the prescriptive rule learned at school in 'simple' cases, in which the distance between the comparative and the comparative marker is small, whereas this rule will not be used when this distance increases. Therefore one of the things annotated is the amount of elements standing in between the comparative and the comparative marker. Whenever the comparative stood directly in front of the comparative marker I annotated a distance of 1, because the comparative is just one position away from the comparative marker, as can be seen in (29a). If just one word occurred between the comparative and the comparative marker the distance annotated is 2, as shown in (29b). For a distance of 3 and 4 see respectively the examples (29c) and (29d).

- (29)
a. gaat iets anders als in Nederland hè
[comp. D: fn007009.37]
'goes a bit different than in the Netherlands eh'
b. Die duurt echt niet langer meer als een uur.
[comp. A: fn000466.47]
'That does not take longer any more than one hour.'
c. Nou, iemand koopt sneller een slaapzak als een tent.
[comp. E: fn000897.240]

- 'Well, a person buys a sleeping bag quicker than a tent.'
- d. Ik heb 't idee dat de microfoon veel dichterbij mij staat dan die bij jou staat. [comp. A: fn000529.73]
 'I have the feeling that the microphone is much closer to me than that it is to you.'

Perhaps the choice for *als* or *dan* in comparative construction is determined by whether we have to deal with a phrasal or clausal comparative. To test whether this plays a role, the complement as well as its type were annotated. The type of complement can be a DP like in example (30a). The sentence in (30b) gives an example of a PP-complement and in (30c) a clausal complement is shown (in this case, the complement is a complementizer phrase, i.e., a CP). Determining the type of complement helps to classify comparative constructions into phrasal and clausal comparatives.

- (30)
- a. Omdat ze vaak een taalachterstand hebben wordt leren ook veel moeilijker dan de Nederlandse taal. [comp. H: fn009161.76]
 'Because they often have a language deficiency learning becomes a lot more difficult than the Dutch language.'
- b. Dus uh ja ik heb 't beter aangepakt dan in Rotterdam. [comp. C: fn008260.134]
 'So, err yes I have tackled it better than in Rotterdam.'
- c. Ik durf te wedden dat jouw ouders makkelijker boeken lezen dan dat jij dat doet. [comp. H: fn009180.13]
 'I bet that your parents read books more easily than you do.'

In annotating the type of complement I restricted myself to the choice between DP, PP and IP/CP. In (31) different DP-complements are used. I considered phrases like (31a) as a determiner phrase because *lezen* 'read' can be regarded as a nominalised verb. Also numerals, like in (31b), are regarded as DPs.

- (31)
- a. Uh ik doe de hele dag niks anders als lezen. [comp. B: fn000130.221]
 'Err I read all day long, I do not do anything else.'
- b. Maar doe alsjeblieft niet meer dan tien. [comp. H: fn009146.15]
 'But please do not do more than ten.'

3.2.3 Random sample of the data

I collected 4565 utterances from the Spoken Dutch Corpus. It was not possible for me to annotate all utterances due to lack of time, so I reduced the amount of data by taking a random sample. In this subsection I will describe the considerations in taking a random sample.

The first thing taken into account was the amount of components. It is very likely to assume that *als* as a comparative marker does not occur in read speech, because it does not occur in texts either. This appeared to be true, since in component o, which only contains read speech, *als* did not occur. Therefore all data of this component were removed. It turned out that the components which only contained prepared non-spontaneous speech did not show any variation in the use of *als* and *dan* as comparative marker. Therefore, besides the component o I also removed the data from the components j up to n. Next, component i did contain just one occurrence of *als* as a comparative marker. Although this component contained spontaneous speech, almost no variation was present. Therefore the data from this component were removed too. By removing these components only the components a until h were left.

With 2929 utterances left, I still needed to reduce the amount of data and decided to halve the data. For this purpose it was important to retain the proportions between the two comparative markers. I divided the data into two groups. The first group only contained occurrences of *dan* and the second group only contained the occurrences of *als*. Within these groups the data were split up into the eight components that were left.

Next I divided the data within the groups and components into the four categories of comparative type, as can be seen in Table 3. Within each cell I also split up the data into two groups: Dutch and Flemish. After dividing the data I deleted half of the data within each cell. This was done at random. The amount of data left is shown in Table 4. After reducing the amount of data only 1465 occurrences were left.

Table 3 On the left: the amount of occurrences of *als* divided by components and comparative types. On the right: the amount of occurrences of *dan* divided by components and comparative types. Before reduction.

<i>Als</i>	ADJ7	ADJ10	VNW24	VNW26	Tot.	<i>Dan</i>	ADJ7	ADJ10	VNW24	VNW26	Tot.
a	12	124	12	68	216	A	50	458	41	267	816
b	1	7	0	9	17	B	14	117	20	96	247
c	1	36	7	31	75	C	13	205	20	133	371
d	1	40	3	19	63	D	3	117	8	73	201
e	0	23	5	9	37	E	0	16	5	11	32
f	2	15	1	7	25	F	21	206	16	186	429
g	0	2	0	2	4	G	13	65	12	88	178
h	2	7	2	5	16	H	9	112	17	64	202
Tot.	19	254	30	150	453	Tot.	123	1296	139	918	2476

Table 4 On the left: the amount of occurrences of *als* divided by components and comparative types. On the right: the amount of occurrences of *dan* divided by components and comparative types. After reduction.

<i>Als</i>	ADJ7	ADJ10	VNW24	VNW26	Tot.	<i>Dan</i>	ADJ7	ADJ10	VNW24	VNW26	Tot.
a	6	62	6	34	108	A	25	229	21	133	408
b	1	4	0	4	9	B	7	59	10	48	124
c	0	18	3	15	36	C	6	102	10	66	184
d	1	20	2	9	32	D	2	58	4	37	101
e	0	11	2	4	17	E	0	8	2	6	16
f	1	8	1	4	14	F	10	103	8	94	215
g	0	1	0	1	2	G	7	33	6	43	89
h	1	4	1	3	9	H	5	56	8	32	101
Tot.	10	128	15	74	227	Tot.	62	648	69	459	1238

3.2.4 Variables for statistical analysis

The linguistic variables that will be used in the statistical analysis are the ones that are described above, although the linguistic variable *complement type* is replaced by the variable *grammatical category*, which contains the levels conjunction and preposition, since a PP-complement only occurs 138 times. Therefore, it is better to take the DP- and PP-complements together. Whenever the annotation of the complement type is a DP or a PP the grammatical category of the comparative marker can be classified as a preposition, as discussed in Section 2. The comparative marker is considered to be a conjunction whenever its complement is a clause.

It is also possible that the choice for *als* or *dan* is influenced by certain properties of the speaker, for instance gender, age and of course education. This kind of information can be obtained from the Spoken Dutch Corpus. In this corpus each speaker, and their properties, can be found by means of the speaker code. I collected information of the speaker about *gender*, *age*, *birth region*, *living region* and *education*.

Gender of course is coded into male and female. I have computed the age by means of the birth year of the speaker and I divided the ages in categories. Category one stands for the age from 0-20. Category two contains all ages between 21 and 40. The third category refers to the ages between 41 and 60 and the last category stands for the people older than 60. I do not expect that within these groups effect are present. That is the reason why I have chosen a categorical division instead of a continuous one.

Another variable that might be of importance is *birth region* and *living region*. As I pointed out in the previous section the use of *als* may be stimulated by dialects in which *als* is used in comparative constructions of inequality. Both variables are coded into five regions. I have divided the Netherlands into north, middle and south. Belgium is divided into a northern region, which stands for Flanders, and a southern one, which stands for

the Walloon provinces. The northern region of the Netherlands consists of the provinces Friesland, Groningen, Drenthe, Overijssel and Noord-Holland. The middle region contains Zuid-Holland, Utrecht, Gelders rivierengebied and the Veluwe up to the river IJssel. The provinces in the south of the Netherlands, which belong to the southern region, are Limburg, Noord-Brabant and Zeeland. In case of a living or birth region outside the Netherlands and Belgium or if it were unknown this was coded as missing, since this corpus study only focuses on Dutch within the Netherlands and Belgium.

The last variable I took from the Spoken Dutch Corpus is *education*. This variable might be of importance too, since I expect that people with a high educational background will apply the prescriptive rule and thus use *dan*, whereas low-educated people will show more variation in the use of *als* and *dan* as comparative marker. The corpus had coded the variable *education* into 4 categories already, so I maintained this coding. The education levels were high, middle and low. The fourth category consists of unknown education levels. The education level *high* refers to people who finished their HBO or university. The middle education level stands for a finished secondary education or MBO. The third education level refers to primary education.

There might be a difference between the use of *als* and *dan* in Flanders and in the Netherlands. That is why I also took the variable *country* into account for this statistical analysis. Table 5 gives an overview of the variables used for statistical analysis.

Table 5 An overview of the independent variables in statistical analysis

Linguistic variables	Categories
Comparative code	ADJ7, ADJ10, VNW24, VNW26
Comparative type	Free, pre-nominal, pre-adjectival
Distance	Range:1-4
Grammatical category	Preposition, conjunction
Non-linguistic variables	Categories
Gender	Man, woman
Age	0-20, 21-40, 41-60, 60+
Birth region	North-N, Middle-N, South-N, North-B, South-B, unknown
Living region	North-N, Middle-N, South-N, North-B, South-B, unknown
Education	High, middle, low, unknown
Country	The Netherlands, Flanders

3.2.5 Statistical analysis

To test whether my expectations are right I will conduct a logistic regression. This form of analysis is able to predict which variables play a role in the choice for either *als* or *dan* as a comparative marker.

In the standard method for regression analysis, linear regression, only continuous variables can be predicted. However, *comparative marker* is not a continuous, but a binary categorical variable. The levels of this variable are *als* and *dan*. Logistic regression is suitable in this case because it is especially developed for a binary categorical dependent variable. Most of the independent variables I will use in my analysis are categorical. An example is *education*. Here four categories are possible, but they cannot be arranged on a scale. The only continuous independent variable I will use in the statistical analysis will be the variable *distance*. I will use the computer program IBM SPSS Statistics 19.0 to carry out this kind of regression analysis. The logistic regression analysis is carried out with the *ENTER* method (Field, 2009). This means all independent variables that turn out to be important are put in the analysis at the same time. To determine the best model I also used the *STEPWISE* method (Field, 2009).

3.3 Results

3.3.1 Crosstabs

Before carrying out the logistic regression I first examined the crosstabs in order to eliminate the independent variables which did not seem to be important in choosing *als* or *dan*. The dependent variable *comparative marker* was put in the rows of the table while the levels of the independent variables were stated in the columns.

I will go through all of these crosstabs and highlight the important findings. It turned out that no effect of *comparative code* was present, as can be seen in Table 6. An ADJ10-comparative occurred 128 times (16.5%) in combination with *als* as a comparative marker, whereas *dan* in this context occurred 648 times (83.5%). The proportions are more or less the same in sentences with other types of comparatives. In ADJ7-comparatives *als* occurs 9 times (12.7%) and *dan* 62 times (87.3%). VNW26-comparatives occur 74 times (13.9%) with *als* and 459 times (86.1%) with *dan*. In VNW24-comparatives *als* is used 16 times (18.8%), whereas *dan* is used 69 times (81.2%).

Table 6 Crosstab in which Comparative Marker and Comparative Code are compared; No interesting effects are present.

		Comparative Code				Total
		ADJ10	ADJ7	VNW26	VNW24	
Comparative Marker	Als	128	9	74	16	227
	Dan	648	62	459	69	1238
Total		776	71	533	85	1465

Table 7 shows that no effect of Comparative type was present. A free comparative in combination with *als* occurred in 15.6% of the cases (n=206), whereas *dan* in this context was used in 84.4% of the cases (n=1114). Almost the same proportions can be seen in pre-nominal comparatives. *Als* was used in 18.0% of the cases (n=16) and *dan* in 82.0% (n=73). Pre-adjectival comparatives hardly go with *als* (n=5; 8.9%). In most cases people used *dan* (n=51; 91.1%). A difference is present between free and pre-nominal comparatives in contrast to pre-adjectival comparatives, but this difference is not big enough to result in an effect.

Table 7 Crosstab in which Comparative Marker and Comparative type are compared; No interesting effects are present.

		Comparative Type			Total
		Free	Pre-nominal	Pre-adjectival	
Comparative Marker	Als	206	16	5	227
	Dan	1114	73	51	1238
Total		1320	89	56	1465

Distance did also not appear to be of influence on the comparative marker, as Table 8 shows. More or less the same proportions are present. A distance of 1 occurred in 15.7% of the cases with *als* (n=150) and in 84.3% of the cases with *dan* (n=807). A distance of 2 in combination with *als* covered 17.6% of the cases (n=55), whereas the same distance in combination with *dan* occurred 82.4% of the time (n=257). A distance of 3 occurred in 12.1% of the cases with *als* (n=17) and in 87.9% of the cases with *dan* (n=124). A distance of 4 in combination with the comparative marker *als* occurred 9.1% of the time (n=5), whereas *dan* in this context occurred in 90.9% of the cases (n=50).

Small differences in the proportions are present, but these differences are not big enough to speak of an effect.

Table 8 Crosstab in which Comparative Marker and Distance are compared; No interesting effects are present.

Comparative Marker * Distance Crosstabulation

		Distance				Total
		1	2	3	4	
Comparative Marker	Als	150	55	17	5	227
	Dan	807	257	124	50	1238
Total		957	312	141	55	1465

Next I examined the crosstab of the linguistic variable *Grammatical category* shown in Table 9. This variable turned out to be not of interest as well. As a preposition, *als* is used in 15.0% of the cases (n=184), in contrast to *dan* that occurred 85.0% of the time (n=1042). As a conjunction, *als* occurred 43 times (18.0%), whereas *dan* occurred 196 times (82.0%). No big differences in proportions are found here.

Table 9 Crosstab in which Comparative Marker and Grammatical Category are compared; No interesting effects are present

Comparative Marker * Grammatical Category Crosstabulation

		Grammatical category		Total
		Preposition	Conjunction	
Comparative Marker	Als	184	43	227
	Dan	1042	196	1238
Total		1226	239	1465

The non-linguistic variable *Gender* did not have any influence on the prediction of *als* or *dan* as comparative marker. The proportions are almost equal at the levels of the variable *Geschlecht*. Men used 15.4% of the time *als* (n=127), whereas women did this 15.5% of the time (n=99). No effect is present here.

Table 10 Crosstab in which Comparative Marker and Gender are compared; No interesting effects are present

Comparative Marker * Gender Crosstabulation

		Gender		Total
		man	Woman	
Comparative Marker	Als	127	99	226
	Dan	696	541	1237
Total		823	640	1463

Age did not seem to be of importance either, as can be seen in Table 11. Here people within the age group of 21-40 used in 15.4% of the cases *als* (n=75), whereas people with an age between 41 and 60 used *als* 14.4% of the time (n=65). People who were older than 60 used in 17.2% of the cases *als* (n=80).

Table 11 Crosstab in which Comparative Marker and Age are compared; No interesting effects are present.

		Age			Total
		21-40	41-60	60+	
Comparative Marker	Als	75	65	80	220
	Dan	411	386	384	1181
Total		486	451	464	1401

Table 12 shows the variable *Birth region* in combination with the dependent variable *Comparative marker*. People born in the northern region of the Netherlands use 25 times *als* (12.3%) and 178 times *dan* (87.7%). Almost the same pattern can be seen for people born in the middle region of the Netherlands. They use 53 times *als* (14.8%) and 306 times *dan* (85.2%). For the people born in the northern region of Belgium also the same pattern is present. Here *als* is used 69 times (13.1%) and *dan* 457 times (86.9%). People born in the southern region of the Netherlands use more *als* than people born in other regions. Here *als* is uttered 51 times (23.2%). *Dan* still is more frequent (n=169; 76.8%). This difference is rather big, but I think this difference will not result in an effect of the variable *Birth region* in the logistic regression.

Table 12 Crosstab in which Comparative Marker and Birth region are compared; No interesting effects are present.

		Birth Region				Total
		North-N	Middle-N	South-N	North-B	
Comparative Marker	Als	25	53	51	69	198
	Dan	178	306	169	457	1110
Total		203	359	220	526	1308

In Table 13 the geographical living areas are compared with the use of *als* and *dan* in comparative constructions. In the northern region of the Netherlands people do not really often use *als* in comparative constructions (n=26; 13.5%) in contrast to *dan* (n=167; 86.5%). For the middle region of the Netherlands the same case applies. *Als* is used 66 times (13.8%) and *dan* 412 times (86.2%). The same pattern can be seen for the northern region of Belgium (Flanders). Most people use *dan* (n=470; 87.2%), whereas *als* is less frequent (n=69; 12.8%). In the southern region of the Netherlands, consisting of the provinces Limburg, Noord-Brabant and Zeeland, however *als* (n=51; 40.2%) is more frequent in contrast to the other regions, although it still occurs less than *dan* (n=76; 59.8%). In the other areas *als* occurs in average in 13% of the time whereas in the southern region of the Netherlands *als* occurs in 40.2% of the time.

Table 13 Crosstab in which Comparative Marker and Living Region are compared; The amount of *als* in the southern region of the Netherlands is remarkable.

		Living Region				Total
		North-N	Middle-N	South-N	North-B	
Comparative Marker	Als	26	66	51	69	212
	Dan	167	412	76	470	1125
Total		193	478	127	539	1337

In Table 14 the dependent variable *comparative marker* and the independent variable *education* are compared. From this crosstab it becomes clear that people with a high education in general use more *dan* (n=1006) than *als* (n=115). In percentages, around 10% uses *als* and 90% uses *dan*. People with a middle education use more *dan* (n=146) than *als* (n=82) too. Here 35.9% of the people use *als* and 64.1% uses *dan*. As can be seen the amount of *als* among people with a middle education is higher compared to people with a high education. People with a low education, in contrast to the people from the other categories, use more *als* (n = 23; 62.1%) than *dan* (n = 14; 37.9%).

Table 14 Crosstab in which Comparative Marker and Education are compared; An interesting contrast can be seen between high and low education.

Comparative Marker * Education Crosstabulation

		Education			Total
		high	middle	Low	
Comparative Marker	Als	115	82	23	220
	Dan	1006	146	14	1166
Total		1121	228	37	1386

From Table 15 it becomes clear that people living in Flanders use less *als* than people living in the Netherlands. In Flanders in 13.2% of the cases *als* is used (n=70), whereas in the Netherlands this percentage is 16.8% (n=157). This difference is too small to speak of an effect of *Country*.

Table 15 Crosstab in which Comparative Marker and Country are compared; No interesting effects are present

Comparative Marker * Country Crosstabulation

		Country		Total
		The Netherlands	Flanders	
Comparative Marker	Als	157	70	227
	Dan	778	460	1238
Total		935	530	1465

3.3.2 Logistic regression

After examining the crosstabs it turned out that the two variables mentioned above, *education* and *living region*, might play a role in predicting the dependent variable *comparative marker*. Therefore I will use these variables as predictors in the logistic regression. The linguistic variables as well as the non-linguistic variables left did not give rise to use them in the regression analysis. *Distance* does not play a role, since the occurrence of *als* at all levels of this variable in average was equal (13%). The same goes for the other linguistic and non-linguistic variables.

When the variables that seemed to be of importance were known in order to predict the comparative marker, a logistic regression was carried out. The independent variables *living region* and *education* were used in this analysis. Table 16 contains the outcomes of the logistic regression. Here it can be seen that people living in the southern region of the Netherlands in contrast to people from the northern region use significantly more *als* in comparative constructions (B = -1.14, p < 0.001). Middle educated people use significantly more often *als* compared to the high educated people (B = -1.46, p < 0.001). Low educated people also use significantly more often *als* than high educated people (B = -2.46, p < 0.001).

Table 16 Outcomes of the logistic regression. Model contains living region and education to predict the comparative marker.

	B	S.E.	Wald	Odds Ratio
Living region ¹			25.36	
Living region(1)	-0.09	0.26	0.10	0.92
Living region(2)	-1.14	0.30	14.50*	0.32
Living region(3)	-0.02	0.26	0.01	0.98
Education ²			97.93	
Education(1)	-1.46	0.18	68.56*	0.23
Education(2)	-2.46	0.37	45.08*	0.09
Constant	2.34	0.23	100.01	10.22

Note: * p < .001. R² = .103 (Cox & Snell), .177 (Nagelkerke).

¹ Living region (1) = Middle-N, Living region (2) = South-N, Living region (3) = North-B, reference: North-N

² Education (1) = Middle, Education (2) = Low, reference: High.

When looking to Table 17, the classification table, it becomes clear that 84.4% of the comparative markers are predicted right by the model including the variables *living region* and *education*, whereas the baseline is 84.0%.

Table 17 Classification table in which the percentage correct predicted comparative markers is shown.

		Comparative Markers as predicted by the model		Percentage Correct
		als	dan	
Comparative Markers as observed in the data	als	43	169	20,3
	dan	38	1079	96,6
Overall Percentage				84,4

To examine whether the model as tested above fits the data the Hosmer and Lemeshow Goodness-of-Fit Test (Hosmer & Lemeshow, 1989) is carried out. When assuming a significance level of 0.05 the test indicates that the model tested above does not fit the data perfectly, $\chi^2(4) = 10.55$, $p = 0.032 < .05$. However, this test must be regarded with caution since in large samples this test might indicate that the model does not fit the data whereas it actually does. Sometimes the test is not that reliable. With regard to the significant p-value of 0.03 in combination with the outcomes of the classification table, which indicates that the model predicts 84.4% correct, it can be said that the model does fit the data well.

3.4 Conclusion

The aim of the corpus study was to search for factors that are of influence on the choice between *als* and *dan* in comparative constructions of inequality.

It turned out that, in line with the expectation, an effect of education was present. People from the middle-educated group use *als* significantly more than high-educated people. Low-educated people use *als* more than middle-educated people. In other words, high-educated people use more *dan* than the middle- and low-educated people. This must be the result of the prescriptive rule that is taught at school.

I also found an effect of living region. People who live in the southern region of the Netherlands used significantly more *als* than people living in other regions of the Netherlands and Belgium. This can be a result of the influence of the southern dialects on

the comparative constructions (Paardekooper, 1970). In *Brabants* *als* is used as a comparative marker in comparative constructions of inequality.

Another remarkable outcome was that no effect of distance between the comparative and its marker *als* or *dan* was present. I did expect to find such an effect, under the assumption that when people consciously apply the prescriptive rule to use *dan* even though their internal grammar would favour *als*, they might 'forget' to apply the rule when the comparative and the marker are far apart.

4 Discussion

In Section 2 I discussed the factors that were in favour of the use of *als* as a comparative marker in comparative constructions of inequality. These factors all had to do with the principles of analogy and economy. In the first place, *als* occurs in comparative constructions of equality. Analogous to this occurrence of *als* it is likely to use it in comparative constructions of inequality too. Also, *als* in comparative constructions could be preferred over *dan*, because the comparative marker in these constructions is a conjunction or a preposition. The marker *als* is a conjunction or a preposition in all other contexts too. *Dan*, however, is a conjunction or preposition only as a comparative marker. In other contexts it belongs to the classes of adverbs and particles. In Section 2 I also argued that dialects in which *als* is used as a comparative marker of inequality could be of influence in using *als* as a comparative marker instead of *dan* in Standard Dutch too. From the corpus study it became clear that this indeed plays a role. An effect of living region is found. The people living in the southern region of the Netherlands use *als* significantly more often than the people living in other regions. In the southern dialects *als* is indeed used as a comparative marker of inequality (Paardekooper, 1970). These three factors favour the use of *als* in comparative constructions of inequality.

Still, *dan* is more frequent in comparatives than *als*. As I argued in the conclusion of Section 2 the use of *dan* might be due to the prescriptive rule learned at school. An effect of education was therefore expected. Low-educated people may not have learned the rule so well, and are expected to use more *als* than high-educated people, therefore. The corpus study clearly revealed that the effect of education is indeed present. High-educated people use significantly more *dan* than low- and middle-educated people. This shows that high-educated people have more successfully learned to apply the prescriptive rule than lower educated people.

Another prediction was that people use *dan* more often when the comparative and the comparative marker are adjacent to each other, based on the assumption that they have learned the prescriptive rule by recognizing it as a chunk, like *groter dan* 'bigger than'. It might become harder for people to apply the rule when the distance between the comparative and the comparative marker increases, because then they will not so easily discern the comparative in the utterance. An effect of distance, however, was not found. The absence of the effect of distance does not mean that people do not apply the prescriptive rule, since an effect of education is present, but it might indicate that people are not aware (anymore) of applying the prescriptive rule. The distance between the comparative and the comparative marker is of no importance, because the rule has become part of people's internal grammar, and they will use *dan* in all comparatives of inequality, independently of the distance between the comparative and the marker. That is, they do not have to think about it anymore, the use of *dan* has become fully automatized. When people do not apply the rule and use the comparative marker *als* they will be often corrected by other people. This indicates that the influence of the prescriptive rule is very big in using *dan* over *als* as a comparative marker of inequality.

The absence of an effect of distance and the presence of an effect of education is interesting. This shows that prescriptive rules are of influence on spontaneous speech. High-educated people learn to use an artificial construction at school and at a certain point in time the use of this construction is fully automatized, even though grammar-internal factors would favour the use of another construction.

5 Conclusion

In this thesis I sought to find an explanation for the fact that sometimes people use *als* as a comparative marker of inequality instead of the prescribed *dan*. It turned out that certain grammar-internal factors would favour the use of *als* in comparative constructions. *Als* is already used as a comparative marker in comparative constructions of equality, so it seems a good candidate for use in comparative constructions of inequality too. *Als* in other contexts belongs to the classes of conjunction and preposition, just like in the context of comparative constructions. *Dan*, however, only is a conjunction or a preposition when in it is a comparative marker, while it is an adverb or particle in all other contexts. The last factor that favours the use of *als* as comparative marker is the influence of the regional dialects in which *als* in comparative constructions of inequality is normal. In order to find out whether these factors indeed favour the use of *als* instead of *dan* in everyday speech and to what degree, I carried out a corpus study on the basis of the Spoken Dutch Corpus.

Dan, however, is still more common as a comparative marker in comparative constructions of inequality. Apparently, the prescriptive rule that is taught at school overrules the various grammar-internal factors that would favour the use of *als*. Because high-educated people use *dan* significantly more often than lower-educated people, and because this is independent from other factors such as the distance between the comparative and the comparative marker, I conclude that for those (high-educated) people, their spontaneous speech has completely absorbed the prescriptive rule, in the sense that the rule to use *dan* in a comparative of inequality has indeed become part of their internal grammar.

6 References

- Bergen, G. van (2010). *Dutch dan in discourse*. Presentation Semantics in the Netherlands (SiN) day 8, November 5, Radboud University Nijmegen
- Field, A. (2009). *Discovering Statistics Using SPSS*. London: SAGE Publications Ltd.
- Haspelmath, M. (1997). *From space to time: Temporal adverbials in the world's languages*. Munich & Newcastle: Lincom Europa.
- Helmantel, M. (2002). *Interactions in the Dutch adpositional domain*. PhD dissertation, Leiden University.
- Hendriks, P. (1995). *Comparatives and categorial grammar*. PhD dissertation, University of Groningen.
- Hoeksema, J. (1983). 'Negative Polarity and the Comparative'. *Natural Language and Linguistic Theory* 1.3, 403-434.
- Hosmer D.W. & Lemeshow S. (1989). *Applied Logistic Regression*. New York: Wiley.
- Nederlandse Taalunie, Taaluniversum. Retrieved at web site:
<http://taaladvies.net/taal/advies/vraag/354/>
- Oostdijk, N. (2003). 'Het Corpus Gesproken Nederlands: Veelzijdig onderzoeksinstrument voor o.a. taalkundig en taal- en spraaktechnologisch onderzoek'. *LINK* 14(1): 3-6.
- Paardekooper, P.C. (1950). 'Als en dan bij vergelijkingen'. *De Nieuwe Taalgids* 43, 160-167.
- Paardekooper, P.C. (1970). 'Groter als'. *De Nieuwe Taalgids* 63, 329-337.
- Pinkham, J. (1985). *The formation of comparative clauses in French and English*. Bloomington: Indiana University Linguistic Club.
- Stroop, J. (2010). *Hun hebben de taal verkwanseld*. Amsterdam: Athenaeum – Polak & Van Gennep.
- Talmy, L. (2000). *Toward a cognitive semantics*. Cambridge, MA: MIT Press 2000.