HOW CLOSELY DOES THE GRAMMAR OF THE DUTCH WORD 'ER' CORRESPOND TO THE USE OF 'ER' FOUND IN DUTCH CORPORA?

Bachelor's Project Thesis

Isa Apallius de Vos, s3239098, i.m.apallius.de.vos@student.rug.nl

Supervisor: S.M. Jones

Abstract: In the context of natural language processing, it is critical to use grammar rules that are accurate and supported by credible sources. In this research, the grammar of the Dutch word *er* is tested to examine whether the literature on *er* corresponds to its use by native speakers. Four uses of *er* are considered: existential *er* (er_X), pronominal *er* (er_P), locative *er* (er_L), and quantitative *er* (er_Q). Three structures with *er* uses not supported by the literature are examined: adjacent *er*s in the midfield, non-adjacent *er*s in the midfield, and er_X in the prefield with er_L or er_P in the midfield. These structures are explored by using two corpora to find sample sentences and to calculate the frequency of these phenomena. From the results it has been concluded that adjacent and non-adjacent *er*s in the midfield do not occur frequently enough to deem them acceptable (the frequency is <0.05% for both structures in both corpora compared to similar supported structures). The structure of er_X with er_L or er_P occurred frequently enough (the frequency is >2.3% for both corpora compared to the acceptable er_X with er_Q structure) to conclude that it needs more research to test its use by native speakers.

1. Introduction

1.1. Background

Natural language processing or NLP can be seen as a subfield of multiple disciplines such as linguistics, computer science, and artificial intelligence. It is concerned with processing human language by using grammars, statistics and/or neural networks with the goal of being able to make systems analyse and generate language in such a way that they can fully understand and respond to humans. Although there are many different ways to process language, this research will focus on the use of formal grammars to do so.

In the context of NLP, a grammar can be defined as a set of rules that describe how sentences, words, and phrases can be formed in a specific language. It is the system a language is based on, and can be quite complex as it needs to include structures that support every correct sentence in a language and should exclude ungrammatical sentences**.** Because language is evolving constantly, grammars need to be looked at and updated regularly to ensure the inclusion of new or different uses of words or phrases.

1.2. The Dutch word *er*

An example of such an ever-evolving word is the Dutch word *er*. Though originally being the unstressed form of the Dutch adverb 'daar' meaning 'there', it currently has four different main functions in the Dutch language (Donaldson, 2008). The existential *er* is usually found at the beginning of a sentence and introduces the verb, often in a similar way as the English 'there' (1a). The pronominal *er* is an obligatory addition to a preposition (1b). This *er* is used instead of the pronouns 'it' and 'them' when referring to something non-human. The locative *er* is used when one refers to a place (1c). This *er* could be replaced by the word 'daar' if one wants to emphasise the location in a sentence. The final *er* is the quantitative *er*, which is used in combination with a numeral or an adverb of quantity (1d). To refer to these different uses of *er*, the following labels will be used: er_X = existential, er_P = pronominal, er_L = locative, and er_Q = quantitative.

(1) a. *Er*$_X$ loopt een man op straat.
There walks a man in the.street
"There is a man walking in the street."

b. Ik kijk *er*$_P$ vaak naar.
I look *there* often at
"I look at it often."

c. Ik ben *er*$_L$ nooit geweest.
I have *there* never been
"I have never been there."

d. Hij heeft *er*$_Q$ drie.
He has *there* three
"He has three of them."

Some argue that there is a fifth use of *er*, namely the use of *er* as the subject of a passive sentence (1e) (Fontein & Pescher-ter Meer, 2004; Voortman, 2005), while others claim that this use is just a subset of the existential function of *er* (Donaldson, 2008).

(1) e. *Er* wordt gedanst.
There is danced
"There is dancing."

In this research, the *er* as the subject of a passive sentence is considered a subset of the existential *er* and not as a different function of *er*. This will be reviewed in the Discussion section.

As with every word or structure in a language, there are certain rules concerning the use of the word *er*. These rules have been the topic of discussion for many researchers, such as Bennis (1986), Grondelaers, Speelman & Carbonez (2001), Neeleman & van de Koot (2006), Donaldson (2008), Grondelaers et al. (2009), and Webelhuth & Bonami (2019). Although there are some differences in opinion on which specific uses and placements of *er* are deemed acceptable, most researchers agree on the basic rules of the use of *er.*

1.3. Syntactic constraints of *er*

In Dutch, a sentence can consist of an indefinite number of main clauses and subordinate clauses, with a minimum of at least one main clause in a sentence. The following structures of topological fields are assumed for Dutch main clauses (2a) and subordinate clauses (2b), as described by Webelhuth & Bonami (2019):

(2) a. prefield – inflected verb – midfield – (other verb(s))

b. complementiser – midfield – verb(s)

The prefield is constrained to a single constituent, while the midfield can contain zero or more constituents. Given these structures, the word *er* is either found in the prefield, the midfield, or both, with a maximum of one *er* per field and two *er*s per clause. Certain uses of *er* are restricted to certain fields depending on their function and possible other explicit or implicit *er*s in the clause.

Single er in a clause

Prefield

There can only be one explicit *er* in the prefield of a main clause, and this has to be the existential *er* (1a). Pronominal, locative, or quantitative *er*s cannot occur on their own in the prefield (3a-c). It is interesting to note that the word *daar*, the stressed form of *er*, is allowed in the prefield while having a locative or pronominal function (3d).

(3) a. * *Er*$_P$ kijk ik vaak naar.
There look I often at
"I look at it often."

b. * *Er*$_L$ woont Jan.
There lives Jan
"Jan lives there."

c. * *Er*$_Q$ zie ik vijf.
There see I five
"I see five of them."

d. *Daar* woont Jan.
There lives Jan
"Jan lives there."

If there is an explicit existential *er* in the prefield, it is possible to have one or more implicit pronominal and locative *er*s, but impossible to have an implicit quantitative *er* (4a-c).

(4) a. *Er*$_{XL}$ wonen veel mensen.
There live many people
"Many people live there."

b. *Er*$_{XLP}$ loopt een weg naartoe.
There walks a road to
"There is a road to it."

c. * *Er*$_{XQ}$ zijn gisteren twee gestolen.
There are yesterday two stolen
"Two (of them) were stolen yesterday."

Midfield
If there are no *er*s in the prefield, every function of *er* can appear in the midfield of a main clause or subordinate clause, as shown in (1b), (1c), (1d), and (5a).

(5) a. Toen kwam *er*$_{X}$ een man.
Then came *there* a man
"Then a man arrived."

It is important to know that the pronominal *er* can appear either earlier in the midfield or immediately before the preposition. If the latter is the case, the *er* and the preposition will be written as a single word. The placement of the *er* does not change the meaning of the sentence, meaning that example (1b) and (5b) hold the same meaning and information.

(5) b. Ik kijk vaak *er*$_{P}$=naar.
I look often *there*=at
"I look at it often."

Other functions of *er* can also appear both at the beginning of the midfield or later on, although *er* at the start of the midfield is more common. Similar to the examples in (4a-b), *er*s in the midfield can represent multiple functions at once (6a-b). It is not possible to have multiple explicit *er*s in the midfield, as shown in (6c).

(6) a. Ze had *er*$_{LQ}$ slechts twee.
She had *there* only two
"She only had two (of them) there."

b. …dat *er*$_{XQQP}$ twee drie uit gehaald hebben.
…that *there* two three out taken have
"… that two of them have taken out three of them." (Webelhuth & Bonami, 2019, ex. 8c)

c. * Ik ben *er*$_{L}$ *er*$_{P}$ gister naartoe gegaan.
I am *there there* yesterday to gone
"I went there yesterday."

Though Neeleman & van de Koot (2006) argue that multiple *er*s in the midfield are possible as long as they are not adjacent (7), others such as Webelhuth & Bonami (2019) disagree and claim that it is only possible to have at most one explicit *er* in the midfield.

(7) (?) …dat hij *er*$_{L}$ zich *er*$_{Q}$ twee heeft
…that he *there* self *there* two has
aangeschaft.
bought
"… that he has bought himself two there."
(Neeleman & van de Koot, 2006, ex. 19a)

Multiple instances of *er* in a clause
As stated before, the prefield and the midfield can both only have one overt *er*, which could imply that an explicit existential *er* in the prefield could co-occur with any *er* in the midfield. However, only the quantitative *er* can be expressed explicitly in the midfield when an explicit existential *er* is already placed in the prefield. The pronominal and locative *er* need to be implicit there (8a-c).

(8) a. *Er*$_{X}$ keken *er*$_{Q}$ drie.
There watched *there* three
"There were three (of them) watching."

b. *Er*$_{XP}$ keken *er*$_{Q}$ drie naar.
There watched *there* three at.
"There were three (of them) watching it."

c. * *Er*$_{X}$ keken *er*$_{Q}$ drie *er*$_{P}$=naar.
There watched *there* three *there*=at
"There were three (of them) watching it."

1.4. Research question

These rules on the usage of *er* would suggest that certain structures with *er* are not possible and thus not likely to occur in the Dutch language. According to the literature, the word *er* is expected to appear either once in the prefield, once in the midfield, or once in both. The use of two consecutive *er*s in a sentence is not supported by the literature, neither is having more than two *er*s in one clause. The use of two *er*s in one field is generally also not allowed, with again the

exception of Neeleman & van de Koot (2006), who support non-adjacent *er*s in the midfield.

These claims need to be tested, thus leading to the following main research question: *"How does the Dutch word* er *occur in different topological fields in written Dutch, and to which extent do these occurrences of* er *correspond to the literature on* er*?"*. This question will be answered by looking at specific occurrences of *er* that are not supported by the literature and answering the following set of subquestions: *"How often do multiple adjacent* ers *occur in the midfield of a clause compared to a single* er *in the midfield?"*, *"How often do multiple non-adjacent* ers *occur in the midfield of a clause compared to a single* er *in the midfield?"*, and *"How often do sentences with an existential* er *in the prefield and an explicit locative or pronominal* er *in the midfield occur compared to sentences with an existential* er *in the prefield and an explicit quantitative* er *in the midfield?"*.

Given the literature and research on er_{s} it is expected that adjacent *er*s will not occur in the midfield in written Dutch. Non-adjacent *er*s in the midfield might be found, depending on whether or not the assumption of Neeleman & van de Koot is correct. Finally, if er_X occurs in the prefield, it is assumed that no er_L or er_P will be found in the midfield.

2. Method

2.1. Corpus Choice

To answer the general research question of how the grammar of the word *er* differs from its practical use by native spearks, two large collections of language data, or corpora, have been used: *Corpus Hedendaags Nederlands* and *SoNaR*.

Corpus Hedendaags Nederlands (CHN) or *Corpus of Contemporary Dutch* in English is a corpus created by the Dutch Language Institute (INT) to monitor contemporary Dutch (Instituut voor Nederlandse Lexicologie, 2014). It includes over 800,000 texts with more than 400,000,000 words, and has a variety of sources, such as blogs, books, and mainly newspapers. The corpus has been automatically tagged with part of speech and lemma and has a built-in search engine which allows users to search for specific words or structures in the database. It includes filters for different variants of Dutch and different text sources.

SoNaR is a similar corpus to CHN, also made available for researchers by the INT (Oostdijk, Reynaert, Hoste, & Schuurman, 2013). With around 2,000,000 documents and more than 500,000,000 words, this corpus has a large collection of sources from online chats and tweets to reports and policy documents. As with CHN, this corpus has been tagged with part of speech and lemma and has a similar search engine, OpenSoNaR, with query possibilities and filters.

These two corpora were chosen for this research because they were both quite extensive, with a large number of different writers and sources. Another reason was the inclusion of a search engine for both corpora, which allowed for many corpus analysis possibilities. Because both search engines supported Corpus Query Language, it was possible to ask the same query to both corpora and get similar results.

2.2. Corpus Query Language

To examine the corpora and their use of *er*, a specific language is used: Corpus Query Language (CQL). CQL was first created by the IMS Corpus WorkBench (Christ, Schulze, Hofmann, & Knig, 1999) and is now also supported by the Lexicom Sketch Engine (Jakubíček, Kilgarriff, McCarthy, & Rychlý, 2010). This language allows the user of a corpus to set conditions for words, making it possible to search for specific occurrences of words or structures. Table 1 shows a few examples of the syntax of CQL.

Table 1: Overview of CQL examples

CQL code	Meaning
[word = "man"]	Search the corpus for every occurrence of the word 'man'.
[lemma = "go"]	Search for every form of the word 'go'. This will find go, goes, went, etc.
[word = "the"] [pos = "ADJ"] [word = "hat"]	Search for the word 'the', followed by an adjective, then followed by the word 'hat'. This will find structures such as 'the nice hat'

In the context of this research, CQL is used to find certain structures surrounding the use of *er*. For example, to find two consecutive *er*s, one could use the following query:

[word = “er”]{2} within <s/>

In this example, “within <s/>” clarifies that the wanted structure is found within a single sentence, while “{2}” indicates that this structure should occur twice in a row.
A more complex query used in this research is:

[word = "er"]
[pos != "VRB.*" & pos != "CONJ.*" & word != "die" & word != "om"]{1,3}
[word="er"]

This query is used to search for non-adjacent *ers* in the midfield. It looks for a single *er*, then one, two, or three words that are not verbs or conjunctions and that are not the words ‘die’ or ‘om’, and finally it looks for another *er*. The full collection of queries used in this research can be found in the Appendix.

2.3. Corpus Settings and Considerations

Both corpora gave the option of selecting certain filters to search for sentence structures from specific sources or language variations. For CHN, no filters were used, while for SoNaR’s search application, OpenSoNaR, a filter was used to exclude the database *The Corpus of Spoken Dutch* or CGN. The exclusion of the CGN database was chosen to ensure that the data did not contain errors related to *er* caused by stutters or a slip of the tongue. The different language variations of Dutch spoken in the Netherlands, Belgium, Suriname, and the Netherlands Antilles were included in this research to allow for a large variation in sources and writers. Although these variations differ in their accent and word use, they were not expected to have different rules or uses of *er*.

Finally, in the context of forms of *er*, the decision was made to exclude the word *d’r*, a synonym of the word *er*, from the searches related to *er*. The word *d’r* can also be used as the feminine possessor *haar* (*her*), which lead to many search results that were not relevant to the grammar of *er*. However, the prepositional *er* that forms a single word with a preposition, such as for example *erbij* (*with it*), *erop* (*on it*), or *erin* (*in it*), was included in the research. This form can cause some confusion for native speakers about whether the *er* and the preposition should be written as a single word or as two separate words. According to Onze Taal (2020), a Dutch language association, the *er* and the preposition should generally be written together as a single word and considered a single word, with a couple of exceptions. This use was not always found in the corpora, but because having a prepositional *er* attached to a preposition only causes confusion about the spelling and does not cause any confusion about the meaning of the word, it was decided to include this form of *er* in the research.

2.4. Data Collection and Analysis

As was described before, to find an answer to the question whether the grammar of *er* differs from the use of *er*, two corpora were used to find certain structures with multiple *ers* in them, using queries in CQL to find those structures in the corpora. To find a certain structure, such as for example adjacent *er*s in the midfield, one or multiple queries were used to encompass the different word orders and word forms that could be used in that structure. Using those queries on the corpora gave a list of sentences that contained the structure that was being examined. Because the number of found sentences was quite large for a number of queries, it would have been difficult to examine and analyse every single sentence. It was thus chosen to take samples of 50 sentences per query and only analyse those sentences. The analysing process included checking whether the sentences found by the query were relevant, labelling the *er*s with a function where possible, and listing other information about the sentences. If the use of an *er* in a sentence could not be determined, no label was added to that *er*. These analyses were then used to give a general overview of the structures in the corpora, what the sentences looked like, and how frequently they were estimated to occur in the corpora. Because the samples were used to estimate the frequency of the examined structures, it should thus be noted that the frequencies found in the Results section are estimations and not exact numbers.

Because this research focuses on the occurrences of sentences that are not or not strongly supported by the literature, the frequency of these sentences is compared to the frequency of similar sentences that are supported by the literature. This comparison is done to give an overview of the occurrences of unsupported *er*s and to find out whether they are relevant in a broader context. In practice this means that sentences with adjacent and non-adjacent ers in the midfield are compared to sentences with a single er in the midfield. Sentences with er$_X$ in the prefield and er$_L$ or er$_P$ in the midfield are compared to sentences with er$_X$ in the prefield and er$_Q$ in the midfield. It was important to give a threshold to the possible acceptability of a structure not supported by the literature. This is why it was decided that the structure should occur in at least 1% of the sentences found for that category. For example, this would mean that for sentences with at least one *er* in the midfield, at least 1% of these sentences should contain two adjacent *er*s for this structure to be seen as a possible new or different use of *er*, as opposed to a use of *er* that is caused only by typing and grammar errors. This 1% was estimated to be a good indicator of whether a phenomenon was caused by chance or by an actual different use of *er*.

3. Results

3.1. General Data Analysis

Table 2: Sentence frequency of single *er* and double *er* sentences per corpus

Er use	Number of sentences CHN		Number of sentences SoNaR	
Sentences with one *er*	2,420,079	95.62%	2,877,238	95.90%
Sentences with two *er*s	110,925	4.38%	123,077	4.10%
TOTAL	2,531,004		3,000,315	

To answer the question of how the grammar of *er* differs from its use by native speakers, three sentence structures not supported by the literature were examined: adjacent *er*s in the midfield, non-adjacent *er*s in the midfield, and *er*$_X$ in the prefield with *er*$_L$ or *er*$_P$ in the midfield. Because of the number of rules regarding the word *er* and its many different uses, it is important to give an overview of the scale of the data and how *er* occurs in it. Table 2 shows the estimated total number of sentences with *er* found in the corpora Corpus Hedendaags Nederlands and SoNaR (2.5M for CHN and 3M for SoNaR), which consists of approximately 96% single *er* sentences and 4% double *er* sentences for both corpora. It should be noted that in this context, a sentence with two *er*s does not necessarily contain two *er*s in a single clause. It refers to complete sentences with two *er*s, including sentences with multiple main or subordinate clauses. The difference between the number of sentences with two *er*s in Table 2 (111K for CHN and 123K for SoNaR) and the number of sentences with two *er*s that will be examined in Tables 3 and 4 (3.4K for CHN and 5.3K for SoNaR) can be explained by the fact that this research focused on sentences that have two *er*s in a single *clause*. It can thus be concluded that the vast majority of the category 'sentences with two *er*s' is comprised of sentences with *er*s in different clauses, and not of sentences with multiple *er*s in the same clause.

3.2. Adjacent and non-adjacent *ers* in the midfield

Table 3: Sentence frequency of sentences with two *er*s or one *er* in the midfield per corpus

Er use	Number of sentences CHN		Number of sentences SoNaR	
Adjacent *er*s in midfield	144	0.01%	210	0.02%
Non-adjacent *er*s in midfield	367	0.03%	590	0.04%
Single *er* in midfield	1,219,394	99.96%	1,401,757	99.94%
TOTAL	1,219,905		1,402,557	

As can be seen in Table 3, the frequency of sentences with adjacent *er*s in the midfield (0.01%) and non-adjacent *er*s in the midfield (0.03%) are both very low compared to the frequency of sentences with a single *er* in the midfield (99.96%) for the *Corpus of Contemporary Dutch* CHN). Similar results were found in the SoNaR corpus with the frequency of sentences with adjacent *er*s

in the midfield (0.02%), non-adjacent *er*s in the midfield (0.04%), and sentences with a single *er* in the midfield (99.94%). The sources for the sentences with adjacent and non-adjacent *er*s were mainly newspapers, although some sentences from SoNaR also originated from autocues, discussion lists and other sources.

After examination of the samples from the dataset, is was found that for adjacent *er*s in the midfield, most of the sentences appeared to be typing errors rather than new or different uses of *er*. An example of a sample sentence from the Corpus Hedendaags Nederlands can be found in example (9).

(9) In 2001 was *er* *er* nog een toename van
In 2001 was *there there* still an increase of
0,8 procent…
0.8 percent
"In 2001 there was still an increase of 0.8 percent…"

In this sentence, it seems that the second *er* does not add any information to the sentence or make it clearer or easier to read.

For non-adjacent *er*s in the midfield it was found that the majority of the sample sentences contained a preposition (10).

(10) Wij denken dat de Europese unie *er* in
We think that the European Union *there* in
deze fase *er* alles *aan* zal doen de
this phase *there* everything *on* will do the
euro te redden.
euro to save
"We believe that the European Union will do everything in its power to save the euro at this stage."

As was explained in the Introduction section, it is possible for a pronominal *er* in the midfield to appear either at the beginning of the midfield or later on in that field. The samples might thus indicate that the people who wrote these sentences with non-adjacent *er*s in the midfield used both of the acceptable placements of *er* in the midfield at the same time. This phenomenon will be examined further in the Discussion section. As with the adjacent *er*s in the midfield, this phenomenon seems more like an error than a new use of *er*.

3.3. Existential *er* in the prefield and locative or pronominal *er* in the midfield

Table 4: Sentence frequency of sentences with *er*$_X$ in the prefield and *er*$_L$/*er*$_P$ or *er*$_Q$ in the midfield per corpus

Er use	Number of sentences CHN		Number of sentences SoNaR	
er$_X$ in prefield with *er*$_L$ or *er*$_P$ in midfield	67	2.32%	107	2.36%
er$_X$ in prefield with *er*$_Q$ in midfield	2821	97.68%	4419	97.64%
TOTAL	2888		4526	

The results in Table 4 show that having *er*$_X$ in the prefield with either *er*$_L$/*er*$_P$ or *er*$_Q$ does not occur often compared to the number of single *er* sentences or even compared to the number of sentences with multiple *er*s found in Table 2. However, in the context of sentences with *er*$_X$ in the prefield and an *er* in the midfield, the combination of *er*$_X$ and *er*$_Q$ occurs more often (97.68% and 97.64% for CHN and SoNaR respectively) than *er*$_X$ and *er*$_L$ or *er*$_P$ (2.32% and 2.36%). The combinations of *er*$_X$ + *er*$_L$ and *er*$_X$ + *er*$_P$ both occurred approximately equally in the sample sets. From the samples from CHN it was found that the combination of *er*$_X$ and *er*$_P$ occurred very frequently before the introduction of a subclause, whereas no such observation was found from the SoNaR samples. This was mainly the case for sentences with *er*$_X$ in the prefield and an *er*$_P$ attached to a preposition. It also seemed that the samples from CHN contained a number of passive sentences, while this was not the case for SoNaR. A sample sentence from the data set is shown in example (11).

(11) *Er* wordt *er* op toegezien dat slechts
There is *there* on supervised that only
consumentenvuurwerk wordt binnengehaald.
consumer.fireworks is brought.in
"It is ensured that only consumer fireworks are brought in."

This sentence is both passive and has an *er* that introduces a subclause. Here, the verb 'worden' indicates a passive sentence and 'erop toezien

dat', meaning 'to ensure that', introduces the subclause. Other patterns concerning the phenomenon of having an *er*X with an *er*L or *er*P could not be found in either corpus. The sources from CHN were mainly newspapers, whereas SoNaR had a variety of sources such as subtitles, newspapers, magazines, discussion boards and autocues.

4. Discussion

4.1. Research Summary

To understand the meaning of the results and in a broader context, it is important to summarise the findings of this research. As was explained in earlier sections, this research focused on the Dutch word *er* and how its use described by the grammar writers of Dutch differs from the use of the word by native speakers. This was done to ensure that programs that use Dutch grammars for processes such as natural language processing have grammars that accurately reflect the use of the Dutch language by native speakers. In this research, there was a focus on structures with *er* that were not supported by the literature to examine whether they would be used by native speakers, despite the literature stating them to be incorrect. The three chosen unsupported structures were sentences with adjacent *er*s in the midfield, sentences with non-adjacent *er*s in the midfield, and sentences with *er*X in the prefield and *er*L or *er*P in the midfield. To check these uses of the word *er*, two corpora, CHN and SoNaR, were chosen to find sentences with these structures in them. They were then compared to similar sentence structures with *er* that were supported by the literature to see how frequently these unsupported structures occurred. It was found that sentences with adjacent and non-adjacent *er*s in the midfield, which were compared to sentences with a single *er* in the midfield, did not occur frequently and consisted mainly of typing errors and grammar errors, rather than new uses of *er*. For sentences with *er*X in the prefield and *er*L or *er*P in the midfield, it was found that, compared to sentences with *er*X in the prefield and *er*Q in the midfield, the structure occurred frequently enough to state that its use is not only the result of typing and grammar errors.

4.2. Why unsupported structures of *er* occur

To explain the results that were found for the different structures of *er*, it is necessary to look at several samples from the dataset. These will explain the contexts in which the sentences with *er* occurred and show what might cause a sentence with multiple *er*s to occur.

For sentences with adjacent *er*s in the midfield, there were a couple of different contexts in which they appeared. One of these is the sentence structure where the second *er* is part of a known phrase in Dutch (12).

(12) Is er een nieuw jong geboren in de
Is there a new young born in the
groep, dan zijn ze *er* *er* als de kippen
group, then are they *there there* as the chickens
bij.
with
"When a new young is born, they'll be there quickly."

In this example, "er als de kippen bij zijn" is a phrase that means to rush to a place or to be somewhere quickly. It might have been the case that the writer of this sentence added the second *er* because the *er* was considered an important part of the phrase that should explicitly be included, instead of just a single second *er* in the sentence. Similarly, there were sample sentences where the second *er* was linked to the verb (13).

(13) Fijn dat je me *er* *er=op* hebt gewezen.
Nice that you me *there there*=on have pointed
"Thank you for pointing it out to me."

Here, the Dutch phrase "erop wijzen", which translates to "pointing out", might have been considered a single phrase where the *er* must always be explicit. In that case, the placement of the second *er* would be explained. Many of these types of structures were found, with verbs such as *ervoor zorgen (to ensure)*, *erin slagen (to succeed in)*, *ervan vinden* (*to have an opinion on)*, and *ervan overtuigd zijn* (*to be convinced of).* As with example (13), the *er*s in this context are pronominal. The other context in which adjacent *er*s in the midfield occurred was shown in example (9) in the Results section. Here, the *er*s are not linked to a verb or phrase, indicating that these sentences are probably caused by typing errors.

For sentences with non-adjacent *er*s in the midfield, it was interesting to note that for native speakers the acceptability of the second *er* seemed higher if there were more words in between the two *er*s. A sentence such as shown in example (10) with three words in between the *er*s could be compared to a sentence with one word in between such as in example (14).

(14) In Zeebrugge zag *er* het *er* rustig uit.
In Zeebrugge looked *there* it *there* quiet out
"Things looked quiet in Zeebrugge."

While the double use of *er* in example (14) appears to be a typing error, the second *er* in example (10) seems more like a grammar error. As was noted in the Results section, many of the sample sentences appeared to contain at least one pronominal *er*. It could be possible that the second *er* in example (10) was explicit because the writer of the sentence forgot that the first *er* was already placed in the sentence. As was stated before, the pronominal *er* is allowed to be placed in different locations in the midfield, which could have led to confusion. However, another theory could be that the second *er* is used in a similar way to a resumptive pronoun, which is a pronoun in a subclause that is used to refer to an antecedent in an earlier clause (15).

(15) *Ik zie de hond die ze *hem* heeft getekend.
I see the dog that she *him* has drawn.
"I see the dog that she has drawn him."

In such a sentence, the word *hem* (*him)* in the subclause refers back to the word *hond* (*dog*) in the main clause. This structure is not supported by the grammar of Dutch or English, but is used in other languages such as Akan to make a sentence easier to interpret (Lartey, 2020). In the context of non-adjacent *er*s in the midfield, it is possible that a second *er* further down the sentence is put there to reemphasise the fact that the *er* is needed in that sentence.

For sentences with er_X in the prefield and er_L or er_P in the midfield, it was interesting to see that certain sentences appeared to be more acceptable to native speakers than others. Two theories about this phenomenon will be explained: the use of *er* as an introduction to a subclause, and the use of *er* in passive sentences.

Firstly, there is a possibility that a sentence with two *er*s is deemed more acceptable grammatically by native speakers if the second *er* introduces a subclause. As was stated in the Results section, it was found that a part of the sample sentences with er_X in the prefield and er_L or er_P in the midfield had structures that did seem to introduce a subclause. These sentences were mainly sentences with an existential *er* in the prefield and a pronominal *er* attached to a preposition in the midfield, as shown in example (11) and in a new sample sentence in example (16).

(16) Er_X is er_P=voor gezorgd dat nabestaanden
There is *there*=for ensured that relatives
zich veilig konden voelen in het land.
themselves safe could feel in the country
"It has been ensured that relatives can feel safe in the country."

In this sentence, the second *er* is part of the phrase "ervoor zorgen dat" (*to ensure that*) and also has the function of introducing the upcoming subclause. This additional role of the pronominal *er* might have made it more natural for that *er* to be explicit instead of implicit.

Secondly, an *er* can function as the subject if a clause does not have one, which is the case for the first *er* in example (16). The *er* in a passive sentence is always paired with the verb *worden* (to *become*) or *zijn* (*to be*). As was stated in the Introduction section, some grammar writers consider the use of *er* as the subject of a passive sentence to be a separate function of *er*, while others state that this *er* is merely a certain use of the existential *er.* From the samples it seems that this use of the word *er* appears more often in the prefield in combination with an *er* in the midfield, and specifically with an er_P in the midfield that introduces a subclause. It might be the case that if the *er* in a passive sentence is a separate use of *er*, this use is allowed to appear in the prefield together with an *er* in the midfield that is not er_Q. This would then reinforce the argument that the use of *er* as a subject of a passive sentence is a separate use of *er* and not part of the existential *er*.

It is interesting to note that the phenomena of the first *er* of a sentence being the passive subject and having a second explicit *er* introduce a subclause occur together quite frequently in the

Corpus Hedendaags Nederland, but not in SoNaR. This difference might be explained by the fact that CHN mainly consists of texts from newspapers, while SoNaR has a large variety of sources, consisting mainly of discussion boards and other sources that might not have been checked on spelling or grammar before being published. This could have had an influence on the difference in uses of *er* found in the two corpora.

4.3. Research Flaws

To understand the significance of the results, it is necessary to reflect on processes such as writing queries in CQL, finding samples from the corpora, and calculating sentence frequencies. These processes were important to the research but not without flaws.

Firstly, the use of Corpus Query Language caused some restrictions on how to find sentence structures in the corpora. For example, there was the option to specify whether the structure that was being searched should occur in a single sentence, but no such option was available for searches within clauses. Although understandable, it made searching for structures with multiple *er*s in the same clause harder because many structures that were found occurred within a single sentence but not within a single clause. This led to queries needing filters to exclude words that indicated the start of a new clause, such as conjunctions.

Secondly, the instances of *er* in the corpora were not labelled per function, meaning that searching for an *er* in the prefield and an *er* in the midfield would lead to finding sentences with any *er* in the prefield and any *er* in the midfield. To find sentences with er_X in the prefield and er_L or er_P in the midfield, the query needed to explicitly exclude sentences that had a numeral or adverb of quantity after the second *er* to ensure that that *er* was not a quantitative *er*. However, it was found that the quantitative *er* can also appear in a sentence without an explicit numeral or adverb of quantity (17).

(17) *Er* zijn *er* die dat leuk vinden.
There are *there* who that fun find
"There are those who like that."

This caused the queries used to find sentences with er_X in the prefield and er_L or er_P in the midfield to include many sentences where the second *er* was quantitative, which made the sample collection a more time-consuming process.

Moreover, the general inaccuracy of the queries led to the estimations of the sentence frequencies not being completely precise. The most accurate estimation would be the number of sentences with adjacent *er*s in the midfield, because the query for that structure did not need many filters to exclude irrelevant sentences. The estimations for the non-adjacent *er*s in the midfield and sentences with er_X in the prefield and er_L or er_P in the midfield should probably have been higher, because those queries needed very specific filters to exclude certain sentences or sentence structures that were not desired. On the other hand, the estimations for the number of sentences with a single *er* in the midfield or with er_X in the prefield and er_Q in the midfield were probably too high because the queries were not specific enough. However, because of the large differences between the scale of the data and the fact that both corpora showed similar results, it was decided that the estimates made in the Results section are sound enough to draw conclusions from the found data.

One way to improve querying the corpora would be to adjust the queries to be more specific, to ensure that the queries do not include as many irrelevant sentences. This takes more time but would ultimately yield more accurate results.

5. Conclusion

5.1. Final research conclusion

To find an answer to the question of whether the rules of the Dutch word *er* correspond to the use of *er* in practice, it was decided to look at certain structures not supported by the literature on *er* to see whether these structures were being used by native Dutch speakers. The examined structures were sentences with adjacent *ers* in the midfield, sentences with non-adjacent *ers* in the midfield, and sentences with er_X in the prefield and er_L or er_P in the midfield. These 'edge cases', structures that may or may not be used by native speakers, were found by using two corpora, CHN and SoNaR. These corpora have been used to find samples of uses of *er* not supported by the literature to find the frequency

of these phenomena and the context in which they occur. From these findings, the following conclusions have been drawn.

Given the samples taken from the corpora and the frequency of this phenomenon, it has been concluded that for sentences with adjacent *er*s in the midfield, the literature was correct to exclude this structure as a correct use of *er*. This structure does not occur frequently enough to deem it a new or acceptable use of *er*. The occurrences of adjacent *er*s in both corpora show that, apart from a small number of understandable grammar mistakes, the majority of the data cannot be explained by any grammar rules or linguistic theories. This majority is thus suspected to consist of typing errors and other mistakes, rather than explicable and possibly new uses of *er*.

From examining the samples of sentences with non-adjacent *er*s in the midfield and looking at the frequency of these sentences, it has been concluded that the literature was correct to exclude this structure as a correct use of *er*. This structure does not occur frequently enough to deem it a new or acceptable use of the word *er*. Although the cause of this phenomenon is unclear, the very low frequency suggests that these sentences are explicable mistakes rather than new or correct uses of *er*. This means that the conclusion on the use of non-adjacent *er*s corresponds with the majority of the literature on *er*, but does not support the claim of Neeleman & van de Koot (2006).

Finally, the frequency and the examined samples of the combination of er_X in the prefield with er_L or er_P in the midfield show that this stucture is a rare occurrence compared to the accepted structure of er_X in the prefield with er_Q in the midfield. However, the question remains whether this relatively low frequency indicates that the phenomenon only consists of errors, or if it is based on a use of *er* not captured by the current grammar theories. As the results suggested, the acceptability of a set of these sentences is unclear. Furthermore, the frequency of this phenomenon, although fairly low, is not believed to be low enough to be able to state that the occurrence of this phenomenon is caused only by typing errors and grammar errors. The frequency and the reaction of native Dutch speakers to some of the samples regarding er_X with er_L or er_P suggest that these sentences might be deemed correct by native speakers in certain circumstances. It can thus be concluded that the literature on the use of explicit *er*s in both the prefield and midfield of a clause might not completely correspond to the usage found in the data from the corpora, and that more research on the acceptability of these sentences should be done.

In conclusion, this research focused on finding an answer to the question whether the grammar of the word *er* corresponds to its use by native speakers. From the gathered data, the analyses, and the answers to the subquestions, it can be concluded that the majority of the grammar rules of *er* corresponds to its use found in the used corpora, with the exeption of the use of *er* in sentences with an *er* in the prefield and an *er* in the midfield of a clause. This research has shown that the grammar rules of *er* are not definitive and should be examined constantly and closely to ensure that its grammar rules corresponds to its use by native speakers.

5.2. Future research

The continuation of research on the word *er* seems crucial to get a complete understanding of what it means, how it is processed, and how native speakers use it. To test the conclusions of this research, it would be interesting to examine the acceptability of sentences with er_X in the prefield and er_L or er_P in the midfield. This could be done in a sentence acceptability judgment test, where native speakers rate the acceptability of different sentences to see whether certain sentence structures are seen as completely wrong, completely correct, or somewhere in between. This test could also be done on sentences with adjacent or non-adjacent *er*s to verify that the conclusions drawn in this research match the judgment of native speakers.

Another idea for future research is to examine other structures related to *er* that are not supported or described by the literature to see whether they are used by native speakers. It could be interesting to look at sentences where the *er* in the prefield is not an existential *er*, or test whether there is a restraint to the number of implicit *er* functions that an explicit *er* can have. In conclusion, there are many possible ways to learn more about the Dutch word *er*.

References

Bennis, H. (1986). *Gaps and dummies.* Dordrecht: Foris.

Christ, O., Schulze, B., Hofmann, A., & Knig, E. (1999). *The IMS Corpus Workbench: Corpus Query Processor (CQP) - User's Manual.* University of Stuttgart, Germany.

Donaldson, B. C. (2008). *Dutch: a comprehensive grammar (2nd ed.).* Abingdon: Routledge.

Fontein, A., & Pescher-ter Meer, A. (2004). *Nederlandse grammatica voor anderstaligen.* NCB.

Genootschap Onze Taal. (2020). *Er / hier / daar / waar + voorzetsel + werkwoord: los of aan elkaar.* https://onzetaal.nl/taaladvies/er-voorzetsel-werkwoord.

Grondelaers, S., Speelman, D., & Carbonez, A. (2001). Regionale variatie in de postverbale distributie van presentatief er. *Neerlandistiek.nl*.

Grondelaers, S., Speelman, D., Drieghe, D., Brysbaert, M., & Geeraerts, D. (2009). Introducing a new entity into discourse: Comprehension and Production Evidence for the Status of Dutch er "there" as a Higher-level Expectancy Monitor. *Acta Psychologica*, 153-160.

Instituut voor Nederlandse Lexicologie. (2014). *Corpus Hedendaags Nederlands.* https://portal.clarin.inl.nl/search/page/search.

Jakubíček, M., Kilgarriff, A., McCarthy, D., & Rychlý, P. (2010). Fast Syntactic Searching in Very Large Corpora for Many Languages. *PACLIC*, 741-747.

Lartey, N. (2020). *A neurolinguistic approach to pronominal resumption in Akan focus constructions.* Groningen: University of Groningen.

Neeleman, A., & van de Koot, H. (2006). Syntactic Haplology. In M. Everaert, & H. van Riemsdijk, *The Blackwell Companion to Syntax, Volume I* (pp. 685-710). London: Blackwell.

Oostdijk, N., Reynaert, M., Hoste, V., & Schuurman, I. (2013). The construction of a 500-million-word reference corpus of contemporary written Dutch. In P. Spyns, & J. Odijk, *Essential Speech and Language Technology for Dutch: Results by the STEVIN-programme* (p. Chapter 13). Springer Verlag.

Voortman, W. (2005). *The Use of Er.* https://www.dutchgrammar.com/_word_docs/Er.pdf.

Webelhuth, G., & Bonami, O. (2019). Syntactic haplology and the Dutchproform "er". *Proceedings of the 26th International Conference on Head-Driven Phrase Structure Grammar* (pp. 100-119). Bucharest: CSLI Publications.

Appendix

An overview of the queries used to collect the data for this research.

Sentences with one or multiple *er*s in the midfield of a clause

Er use	Query CHN	Query SoNaR
Adjacent *er*s midfield	[word = "er"] [word = "er"] within <s/> + [word = "er"] [word = "er.*" & pos = "ADV.*" & word != "ergens"] within <s/>	[word = "er"] [word = "er"] within <s/> + [word = "er"] [word = "er.*" & pos = "BW.*" & word != "ergens"] within <s/>
Non-adjacent *er*s midfield	[word = "er"] [pos != "VRB.*" & pos != "CONJ.*" & word != "die" & word != "om"]{1,3} [word="er"] within <s/> + [word = "er"] [pos != "VRB.*" & pos != "CONJ.*" & word != "die" & word != "om"]{1,3} [word = "er.*" & pos = "ADV.*" & word != "ergens"] within <s/>	[word = "er"] [pos != "WW.*" & pos != "VG.*" & pos != "LET.*" & word != "die" & word != "om"]{1,3} [word="er"] within <s/> + [word = "er"] [pos != "WW.*" & pos != "VG.*" & pos != "LET.*" & word != "die" & word != "om"]{1,3} [word = "er.*" & pos = "BW.*" & word != "ergens"] within <s/>
Single *er* in midfield	<s> [word != 'er'] []{1,5} [word = "er"] within <s/> + <s> [word != 'er'] [word = "er"] within <s/> + <s> [word != 'er'] []{0,5} [word = "er.*" & pos = "ADV.*" & word != "ergens"] within <s/>	<s> [word != 'er'] []{1,5} [word = "er"] within <s/> + <s> [word != 'er'] [word = "er"] within <s/> + <s> [word != 'er'] []{0,5} [word = "er.*" & pos = "BW.*" & word != "ergens"] within <s/>

Sentences with an *er* in the prefield and an *er* in the midfield of a clause

Er use	Query CHN	Query SoNaR
*er*X in prefield + *er*L or *er*P in midfield	<s> [word = "er"] [] [word = "er"] [pos != "NUM.*" & pos != "PD.*"]{3} within <s/> + <s> [word = "er"] [] [word = "er.*" & pos = "ADV.*" & word != "ergens"] within <s/> + <s> [word = "er"] [pos = "VRB.*"] [pos != "VRB.*" & pos != "CONJ.*" & word != "die" & word != "om"]{1,3} [word = "er.*" & (pos = "ADV.*" | pos = "PD.*") & word != "ergens"] [pos != "NUM.*" & pos != "PD.*"]{3} within <s/>	<s> [word = "er"] [] [word = "er"] [pos_pdtype!="grad" & pos_head != "tw" & word != "?" & word != "eentje"]{3} within <s/> + <s> [word = "er"] []{1,3} [word = "er.*" & pos = "BW.*" & word != "ergens"] within <s/>
*er*X in prefield + *er*Q in midfield	<s> [word = "er"] [pos = "VRB.*"] [word = "er"] within <s/>	<s> [word = "er"] [] [word = "er"] within <s/>

Sentences with one or more *er*s

Er use	**Query CHN**	**Query SoNaR**
Sentences with one *er*	[word = "er"]	[word = "er"]
Sentences with two *er*s	[word = "er.*" & (pos = "ADV.*" | pos = "PD.*") & word != "ergens"] []{0,10} [word = "er.*" & (pos = "ADV.*" | pos = "PD.*") & word != "ergens"] within <s/>	[word = "er.*" & (pos = "BW.*" | pos = "VNW.*") & word != "ergens"] []{0,10} [word = "er.*" & (pos = "BW.*" | pos = "VNW.*") & word != "ergens"] within <s/>