Automatic Content Analysis of Dutch Traffic Accident News

Viability of computational linguistic approaches for discursive content analysis

Jonas H. van Oenen
10670947

Bachelor thesis
Credits: 18 EC

Bachelor Opleiding Kunstmatige Intelligentie

University of Amsterdam
Faculty of Science
Science Park 904
1098 XH Amsterdam

Supervisor
dr. M.J. Marx

Informatics Institute
Faculty of Science
University of Amsterdam
Science Park 904
1098 XH Amsterdam

Feb 1st, 2019

Abstract

Traffic accidents are one of the highest causes of death in the Netherlands, however, widespread debate and discussion on road safety remains absent. Journalist and researchers have hypothesised that this is partly because of discursive practices in the language of new articles on traffic accidents. Discursive content analysis on this subject matter has all been done manually. The goal of this thesis is research the viability to automate this discursive content analysis using computational linguistics technologies. A literary study was done of possible computational approaches for this problem, determining the approach with the most potential. Exploratory pattern analysis was done on the Dutch traffic accident domain to determine if there were patterns which can be identified through computational means. The results show that there are demonstrable patterns which can be found using computational techniques. Furthermore, an unsupervised information extraction approach was determined to be most promising for this problem. In conclusion, it is argued that a computational approach to automating discursive content analysis is possible but has some obstacles to overcome. And that the exploratory analysis shows that patterns are present in Dutch traffic accident news and could be found through computational means.

Contents

1 Introduction

There is not a day that goes by without a tragic headline in the news involving a traffic accident. Most of these traffic accident articles appear on news websites, where news organisation frequently delineate a special category for them.[1] Turning to official statistics, traffic accidents are one of the highest (non natural) causes of death in the Netherlands.[2] Furthermore, the number of heavily injured people due to traffic accidents has been steadily increasing over the past couple of years –while the number of death has been relatively consistent.[3] These statistics are alarming and even more so, their accuracy has been questioned with researchers suggesting that the actual number of victims of traffic accidents are much higher (Kemler & den Hertog, 2009). The government in the Netherlands has tried to address these problems but large scale action remains limited (Aarts, Eenink, Weijermans, Knapper, & Schagen, 2014).[4] This all begs the question: why is it that these issues are not more largely debated and protested?

Recently, some journalists have been raising questions and awareness about traffic accidents(Verkade & Brömmelstroet, 2018). Starting a debate on the way we see road safety and more specifically: debating the ways language influences our portrayal of road safety (Verkade, 2017). Language has a profound influence on the way we understand things and is therefore a powerful tool in the framing of information. News articles on traffic accidents have a very distinct style of reporting, characterised by a factual and seemingly neutral description of the events. Research has been done on the discourse in news reporting on traffic accidents, trying to find these distinctive features in the language through discursive content analysis (Ralph, Lacobucci, Thigpen, & Goddard, 2019). This content analysis has mostly relied on manually analysing articles or simple statistical analysis of a small amount of articles. This thesis hypothesises that Dutch traffic accidents news has specific patterns and that these can be detected through computational linguistic techniques. The central question is therefore: *how can computational linguistics techniques be used to find patterns between entities within traffic accident news?* The patterns and entities mentioned can be understood through an example as "car hits cyclist", where 'car' and 'cyclist' are entities and their relation or pattern is 'hits'.

In the first section, the concepts of discourse and framing will be explained together with research showing the real impact language has on the public image of certain problems. These concepts form the basis on which most content analysis on traffic accidents has been done. An overview is given of the discursive content research that has been done on road safety. A central role is

[1] For example: `https://www.nu.nl/tag/verkeersongevallen`, `https://www.parool.nl/alle-nieuws-over-verkeersongevallen/`, `https://www.volkskrant.nl/alle-nieuws-over-verkeersongevallen/`

[2] `https://www.volksgezondheidenzorg.info/ranglijst/ranglijst-doodsoorzaken-op-basis-van-sterfte`

[3] `https://www.cbs.nl/nl-nl/maatschappij/verkeer-en-vervoer/transport-en-mobiliteit/mobiliteit/verkeersongevallen/categorie-verkeersongevallen/doden-en-gewonden-in-het-wegverkeer`

[4] The most recent attempt can be found here: `https://www.verkeersveiligheid2030.nl/default.aspx`

given to the research of Ralph et al. (2019), elaborating on their approach to the problem of discursive content analysis of traffic accidents . The second section sets forth the possible computational approaches and state-of-the-art algorithms applicable to this problem. Specifically detailing the concept of unsupervised relation extraction as the the approach with the most potential for finding patterns in unstructured data like news articles. Having discussed the current methods in content analysis and computational approaches which could automate this process, the third section will show the method this research has taken for preliminary work on the subject. Showing that there are indeed patterns to be found in the Dutch news on traffic accidents, which could in theory be found using more automated computational approaches. Concluding with the overall potential and possibilities for automatically finding patterns between entities which are involved in the traffic accident. Finally, the shortcomings and limitations of this research will be discussed.

2 Theoretical context

Why is it that most often it is a 'car' that hits someone and not a 'driver'? Why is it that it is never a 'bicycle' that hits somebody but a 'cyclist'? To answer these questions it is important to look at the role that language plays in how we see things. Research on language use, what it means and what it represents, cannot forego some basic knowledge of the way in which language is formed. The first section will answer the question how language is formed by power and the role discursive practices have, explaining what discursive content is. The second section will discuss the research which has been done on discursive content analysis traffic accident news. Firstly, there have been multiple research projects in Belgium on the reliability of news reporting on traffic accidents using governmental databases. A similar research project was considered in Dutch context for this thesis but limited data made this impossible as will be explained. Secondly, numerous research has been done in discursive content analysis of traffic accident reporting. This research builds upon the ideas of discourse and framing, manually analysing news articles and finding patterns and language use corresponding to a specific perspective of traffic accidents in the general public. Specifically, the recent research of Ralph et al. will be discussed more in depth –looking at how they framed their analysis and results (Ralph et al., 2019).

2.1 Discursive practices

What are discursive practices? These are practices which are the consequence of the discourse in a domain. The concept of discourse, originally used by Foucault, is that through which a subject is constituted and generates knowledge (Xie, 2018). It entails far more than this short description. However, in the context of this paper one aspect of this idea of discourse is of great importance. The feature of language as a means through which knowledge on the subject is instantiated. Language cannot be interpreted as a objective means through which information is simply communicated. Language itself forms the knowledge and the information of a subject (Xie, 2018, p.400). The terms, concepts and descriptions used in a specific context are paramount to the way we under-

stand a subject. This idea of language as forming knowledge also gives rise to a complication, the intertwinement with power. Stating that language is forming for knowledge also means that language can be used to alter the discussion simply by changing the way we talk about a subject. The power to define the language which is used to discuss a specific subject therefore becomes important. These power relations are not simply on display for everyone to see, they are simply present in discursive practices. They are deeply embedded in the concepts and terms we use, with no real perpetrator to be singled out. The task in discursive analysis is to uncover these relations, not too appoint blame, but to make clear the limitations and biases in the language used in a domain.

An important aspect of discursive practices in news is the concept of framing. The concept of framing is defined as the context in which a news article is presented (Scheufele & Tewksbury, 2006). And more specifically, the way this context refers to knowledge and biases already present in the target audience of the article. The core idea is that there is a specific way in which a subject is presented affecting how it will be read. In the context of the traffic accident domain, one could take the example of an article of a car hitting someone and the article avoiding to blame the car since most people driving cars will feel persecuted. There has been research on the kind of framing in broad traffic news, arguing that news has an important impact on the public image of traffic accidents(Mejia, 2017). The report by Meija looked at news articles in the San Francisco and concluded that articles are usually framed as a conflict between individuals. Furthermore, That there were strict divisions between road users like cyclist and car drivers and that these divisions were framed as a danger to each other. However, news articles usually ignored the fact that road safety was a broader issue and that –for example– unsafe road strategies are also at fault.

2.2 Discursive research in traffic accident news

Central to this thesis is research into discursive content analysis of traffic accident news. What research has been done in discursive content analysis of traffic accidents? Identifying these discursive practices in traffic accident news has had two main approaches.

1. Identifying discursive practices by analysing news articles and detecting fixed patterns in the language that is used in reporting.

2. Identifying discursive practices by relating news articles to official governmental databases. The features of the official databases are compared with information in news articles to see what information is deemed most important in news.

2.2.1 Identifying discursive practices through news articles

Content analysis in South Africa identified dominant and non-dominant factors in national traffic accident news (MacRitchie & Seedat, 2008). Showing that news specifically framed neglect of individual drivers (speeding, drunkenness) as the main cause of accidents, ignoring "societal, institutional and corporate responsibility in road safety" (MacRitchie & Seedat, 2008, p.350). They also found that reporters tended to pick out empirical data which enforced the stereotype

of individual drivers as the problem. More recent research in Canada identified similar factors, concluding that larger issues in road safety were mostly ignored (Magusin, 2017). However, a key difference is that MacRitchie identified that blame was directed at individual drivers whereas Magusin concluded that drivers were overwhelmingly absolved from blame. Furthermore, Magusin argues that vehicles –specifically cars– are dominant in traffic discourse and usually prioritise vehicle traffic over other road users. In addition, Magusin found that news reporting was "factual and dehumanizing" and that only a small number of news articles humanised the victims and acknowledged broader issues with road safety (Magusin, 2017, p.89). Recent research by Ralph et al. (2019) concurs with the findings by Magusin, arguing that traffic accidents are usually framed as isolated incidents. However, the research by Magusin mostly looked at active versus passive sentences and terms used in articles.

Ralph et al. developed a method to specifically analyse blame in traffic accident articles, building a clear framework in which this can be researched. The data consisted of a small set (200) of newspaper articles scraped from the web, with a specific set of terms present in these articles(for example bicyclist, bike etc.). Half of the articles consisted of someone getting hit by a car while biking, the other half getting hit while walking. The news articles were then analysed for blame using three features: focus, agency and objects versus person based language. The two entities involved with an accident were categorised as either the vehicle or a vulnerable road user (VRU). Using this, each article was as being agentive or non-agentive where for example agentive means that a "A car hit a VRU" and non-agentive "A car and a VRU collided". Then the focus of the sentence was determined i.e. what the central subject of a sentence was. Thirdly, the object versus person language was analysed by identifying if the vehicle was referred to as a driver or vehicle. These three features were identified for each article and the results showed that 65% of sentences had an agent, 75% focused on the VRU and the vehicle was referred to as a vehicle and not driver in 81% of sentences. Ralph et al. argue that these statistics show that VRU are consistently blamed for accidents and specifically drivers of cars are absolved of blame. Concluding that this does not mean drivers of vehicles should be blamed more often but that car accidents should be framed as a public health concern –replacing terms like 'accident' with 'crash' or 'collision'.

2.2.2 Identifying discursive practices through news in relation to official databases

There has been substantial Belgian research into the validity and completeness of traffic accident news in reference to governmental databases (De Ceunynck, De Smedt, Daniels, Wouters, & Baets, 2015; Daniels, Brijs, & Keunen, 2010). Research by De Ceunynck et al. (2015) suggest that there are a number of biases in the kind of traffic accidents that get reported in the news. The severity of traffic accidents had a significant impact on the probability of a traffic accident being reported, with more grievous accidents being reported more often. De Ceunynck et al. suggest that these biases create a serious risk in creating and maintaining skewed perceptions about traffic safety. The method in both these researches was to use an official governmental database with recorded accidents and match these with traffic accident news articles. This has the advantage of

having a reliable measure as to evaluate the results, because the matched articles can be contrasted with accidents that were not reported. Furthermore, it is possible to specifically search for which features are represented in a news article because the governmental databases store numerous conditions of an accident.

To recreate this research in a Dutch context would require a similar governmental database. This database does exist and is called "Bestand Registreerde Ongevallen Nederland" (BRON), it contains individual traffic accidents and a number of features describing the accident. However, there are two major drawbacks to BRON: all features needed to match with articles have been removed in connection with EU privacy legislation (Rijkswaterstaat, 2018); and BRON has significant shortcomings in the number of accidents actually registered and registered accident are often incomplete (Houwing, 2017). These drawbacks make it impossible to recreate the research that was done in Belgium.

3 Computational approaches for discursive research

The goal of this research is to automatically recreate the content analysis as described in the previous subsections. In order to gain an understanding of this problem in computational terms, it is important to position this research in respect to the broader fields of information extraction and natural language processing. The core problem being discussed is how to computationally identify certain linguistic features in the data. The question this section aims to answer is: what (state-of-the-art) computational linguistic approaches are most applicable for this problem?

As for the problem in this thesis: the data is unstructured, unlabelled and specific to the domain of traffic accidents. The field of AI associated with this kind of research is called Text Mining, where the goal is to develop ways to computationally discover knowledge from texts. (Allahyari et al., 2017). Allayhrai et al. divide the field into numerous sub divisions but this thesis is focused on two main approaches: Information Extraction (IE) and Natural Language Proccesing (NLP). The goal of NLP is defined as "understanding natural language using computers" (Allahyari et al., 2017). The data in this research consists of news articles meaning that the main analysis will have to consist of processing raw text, making NLP a logical choice of direction. The goal of IE is described as "automatically extracting information or facts from unstructured or semi-structured documents"(Allahyari et al., 2017). This goal is similar to the goal of content analysis in general, only differing in the fact that IE tries to automate this process. NLP and IE complement each other, since algorithms for IE usually depend on various NLP techniques. An overview is given of both natural language processing and information extraction techniques and recent research in these fields. The current approaches and their relevance to the specific problem and domain of this thesis will be examined, to determine which approach has the most potential.

3.1 Natural Language Processing

NLP is a broad field but for the purposes of this research the lexical and syntactic processing of language is key (Liddy, 2001). The lexical level of processing natural language concerns itself with understanding individual words in sentences, with one of the most widely used techniques being part-of-speech (POS) tagging. At the syntactic level the interest is in the grammatical structure of a sentence, where algorithms which classify these grammatical structures are referred to as dependency parsers. These techniques have been implemented in a variety of programming tool sets like NLTK, Spacy, FROG and Alpino. Alpino and FROG are parsers developed specifically for the Dutch language which implement POS tagging, morphological analysis and dependency parsing (Van den Bosch, Busser, Daelemans, & Canisius, 2007) .[5] POS tagging and Dependency parsing are used in this research to find the underlying patterns in the language –which are the basic building blocks for most IE research.

3.2 Information Extraction

Research in IE can be divided in two main tasks: Named Entity Recognition (NER) and Relation Extraction (Jiang, 2012).

Named Entity Recognition (NER) The task of NER is aimed at classifying a word or sequence of words into a predefined category, examples of these general categories are person, organisation and location. These NER algorithms are used in almost all IE research and is the starting point for many techniques in the relation extraction task. A Dutch NER algorithm has also been implemented which categorises entities into six predefined categories [Desmet, 2013]. These pre-existing NER algorithms work fairly well on some data sets but still only provide categorising in general terms like person and organisation. Specific domains –like traffic accident reporting– require distinct categories like vehicles which ordinary NER algorithms do not provide. In this case, one can train a NER algorithm on a training set to learn a different set of categories. The benefit of this approach is that categorises can be more specific than the generic ones, the disadvantage is that this does require a labelled training set which is not present in many cases. For the purposes of this research NER poses a challenge since the target categories are specific but there is no labelled data. This means that named entities need to be avoided which has significant consequences for the kind of approaches which are applicable.

Relation Extraction The second fundamental task in IE is Relation Extraction, which aims to find the semantic relations between entities in some form of raw text (Jiang, 2012, p. 22). There are three main approaches in Relation Extraction: supervised learning, weakly supervised learning and unsupervised learning.

1. Supervised Relation Extraction classifies relations between entities into predefined categories. The benefit with supervised learning is that the results are precise but the disadvantage is that there needs to be some way of pre-defining categories. In a data set with a large variety in language use and form it is

[5] `http://www.let.rug.nl/vannoord/alp/Alpino/`

difficult to obtain categories which will accurately cover all relations. As will be discussed in the method and approach of this thesis, the variety in our data is large and manually defining categories is difficult.

2. The second approach is called weakly supervised learning, where the goal is to use a small amount of training data to create a model. There are two techniques widely used in this field, the first is called *bootstrapping*. The main idea behind bootstrapping is that a small number of entity pairs which define the target relation are inputted into an algorithm (Jiang, 2012, p. 29). This algorithm then finds the occurrences of these pairs and collects the patterns in which they occur. Then these patterns are retrieved and used to find more entity pairs and vice versa which ultimately culminates in a collection of patterns and the entities with which these occur.

Some good examples of bootstrapping algorithms are Snowball and TIER (Huang & Riloff, 2012; Agichtein & Gravano, 2000). The problem with both these algorithms is that they still depend on NER to find entity pairs which match with patterns. The second technique is called distant supervision, the key idea is the same as bootstrapping but instead of a small amount of seed pairs a database on the web is used (for example wikipedia) (Jiang, 2012, p. 30).. The advantage of using a large database on the web is that the amount of entities which can be detected increases significantly. Research using this technique has been recently implemented and works well on data which corresponds relevantly to Wikipedia entities (Sorokin & Gurevych, 2017). However, the same problem remains that in order for it to work the database needs to be labelled using some sort of NER algorithm.

3. The last approach is called unsupervised information extraction, where the goal is to discover relations without any sort of predefined category. This means that this approach works well for problems where the target relations are not known and there is a considerable amount of differing relations. There are two prominent techniques in unsupervised information extraction: Relation Discovery and Template Induction (Jiang, 2012). Relation discovery aims to find all interesting relations in a text, without the use of a NER making it far more versatile. Shinyama and Sekine implemented an algorithm where the example is given of finding the relations between natural disasters and cities, trying to determine relations as X was hit by Y (Shinyama & Sekine, 2006). This research is similar to the goals of this thesis, trying to determine the relationship X was hit by Y only then in terms of traffic accidents.

A more general approach to unsupervised relation discovery was established by Rosenfeld and Feldman, explicating the use of clustering in their approach (Rosenfeld & Feldman, 2007). This general approach has three main stages, firstly a set of entities is created usually containing pairs. Secondly, sentences with these entities co-occurring are found and a similarity measure between different sentences is used. Lastly, through calculating this similarity measure between sentences clusters are formed of patterns which have a strong resemblance. In the paper by Rosenfeld and Feldman a NER algorithm is explicitly avoided and other means of generating pairs of entities are considered.

The second unsupervised IE technique is called Template Induction. Template induction is similar to relation discovery, Chambers and Jurafsky developed an algorithm which not only finds single binary relations between entities but can infer entire templates from a corpus (Chambers & Jurafsky, 2011). This means that the algorithm does not only find a relation X hit Y but also more complex relations as for instance X hit Y with Z. In theory, relations as victim and perpetrator and the vehicle which caused an accident could be identified using this approach. Furthermore, the algorithm does not rely on named entities for patterns to be learned –making it useful in domain-specific data sets.

In conclusion, there are a few fundamental problems which make it too difficult to apply techniques in IE to the traffic accident domain. NER algorithms are essential to supervised and weakly supervised approaches but existing NER algorithms only consider general categories making them ineffective for specific domains like traffic accidents. Training a NER algorithm on a specific domain requires a labelled data set which is not present in many cases and labour intensive to manually create. Unsupervised learning approaches avoid the use of NER algorithms and research has shown that these can be effective in detecting relations of differing complexity. However, research using these techniques has only involved English data sets and involves numerous English resources. The research by Chambers and Jurafsky has potential for the aims of this thesis since it is able to identify multiple relations and their code is also publicly available.[6] In order to reproduce their work in a Dutch setting a Dutch wordnet would need to be provided, dependency parsers would need to be replaced with Dutch applicable ones and configuration files edited for the specific domain. These challenges are interesting for future research but this thesis aims to provide preliminary work to show that this work would be relevant for the domain.

4 Exploratory pattern analysis in traffic accident news

This section elaborates on the method used in this thesis to produce some preliminary results on the automation of discursive content analysis. Describing the data and constituting a method through which potential patterns were identified. The first subsection will explain how data was gathered and pre-proccessing that was done on the data. The second subsection will give an analysis of the composition of the data. The third subsection will elaborate on the method for exploratory pattern analysis using this data.

4.1 Data: Lexis Nexis

The data for this research is collected through the Lexis Nexis service.[7] This service provides access to articles of a large number of news outlets around the world. Furthermore, it provides the ability to search for articles matching a specific search term and allows the user to set the scope of the search criteria. For the purposes of this research only news articles from Dutch news outlets

[6] `https://github.com/nchambers/schemas`
[7] `https://academic.lexisnexis.nl/`

were searched. A collection of search terms were gathered which generally correspond to news about traffic accidents and give a reasonable representation of the variety of articles. In table 1 the search terms are shown with their respective amounts of articles in the data set.

As the table shows, 6 search queries were used and the amount of articles they yielded differs between 400 and 600. The total number of articles downloaded is approximately 3000, however, this does not account for duplicates. Duplicates were identified based on the exact matching of the text feature. This does mean that –in theory– there could be articles describing the same accidents with slight variation in the text. After removing duplicates the data set consists of 2457 unique articles. Shortly summarising the features: title is the title of an article; text is the main body of text of an article; date is the original date of publication; journal the news outlet it originates from; set label is the search query through which the respective article was found. The title and text feature are the most significant features for the research in this thesis. In the title and text features non-alpha words were removed and all words were lower cased.

	Number of articles
geschept	487
ongeval	479
botsing	469
aanrijding	360
ongeluk	345
aangereden	317

Table 1: Lexis Nexis articles per search term

4.2 Analysis: data composition

The composition of the data is analysed to give some insight into the features and substantiate choices made in the method subsection. Firstly, the dates of articles were gathered to confirm that our data set is relatively current and does not contain articles from decades ago. Language changes over periods of time and using old articles could skew results, even though it would be interesting to see if there are big differences in language patterns over the years. However, for the data analysis to be relevant of language as it is now the articles need to be relatively recent. The collected articles were plotted per year as can be seen in figure 1. The data set consist of mainly articles in the year 2018 and 2019. This is due to the fact that Lexis Nexis prioritises more recent results.

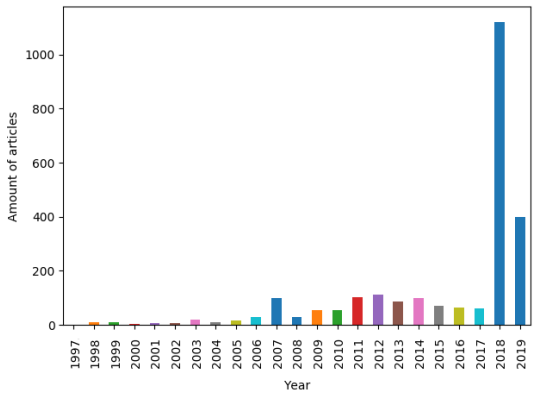


Figure 1: Amount of articles per year

Every article in the data is also annotated with the journal it was originally issued in. This annotation was not always accurate as will be discussed in the discussion section. However, there are still some inferences we can make about the distribution of journals. In order for the pattern analysis not to be biased to the language use in a specific journal, there needs to be some variety in the journals. The sums of articles that journals contributed were calculated as can be seen in figure 2. This shows that a significant part of journals only occurred

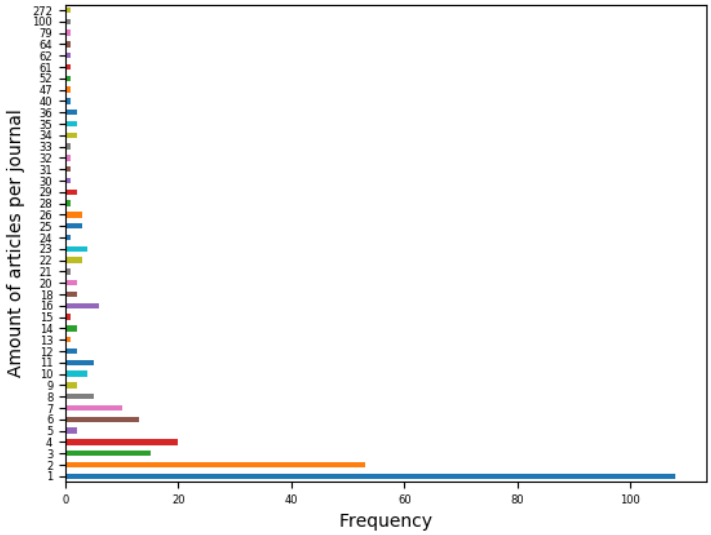
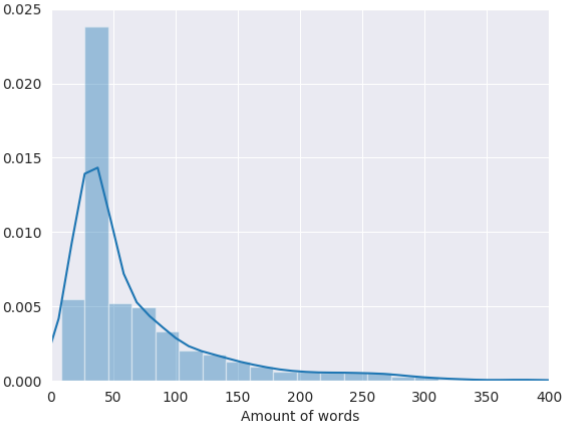


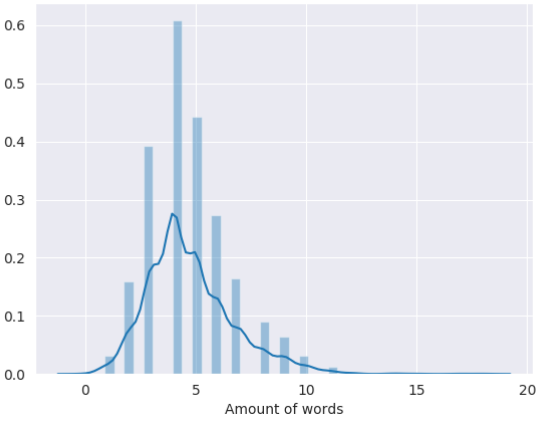
Figure 2: Article sums for all journals

once in the data set. A small number of journals contributed large amounts of articles, for example: one journal contributed 272 articles to the data set (Algemeen Dagblad). Most importantly, this shows that the variety of journals is large which makes it unlikely that the results are biased towards a specific journal. In addition, it also shows that traffic accidents are reported in a wide variety of media with no single journal having a monopoly on the reporting of traffic accidents.

The titles and text of an article are the main target of pattern analysis, statistics show important properties of these features. The amount of words in titles and text were plotted, as can be seen in figure 3. The first plot (a) shows the amount of words in the text of articles. The plot shows that there is a substantial peak around 40 words. There are two explanations for this: Lexis Nexis does not show the full body of an article but usually cuts the article after 40/50 words; and traffic accidents articles are usually quite short so even full length articles might not have more then about 200 words. This does mean that analysis of the text – in some cases – might not correspond to the entirety of the article body.



(a) Histogram and KDE of text



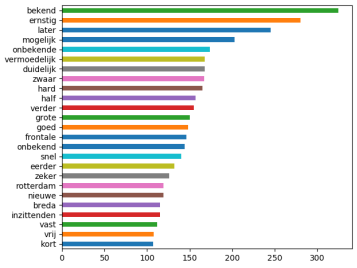
(b) Histogram and KDE of titles

Figure 3: KDE plots of word density in titles and text

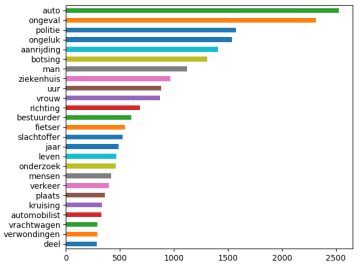
The second plot (b) in figure 3 shows the amount of words in titles. As expected, title are short and usually do not span more then 10 words. However, it is remarkable that there is such a high peak around 4 words. Apparently –right around– 4 words is an optimal amount to on the one hand explain what the article is about and on the other hand keep the attention of the reader. Due to the short title, it is probable that the words contain the core subject of the article and likely those involved. Moreover, in general, news articles titles represent the most important subject of the article.

Two techniques are used in the pattern analysis – also mentioned in the computational approaches section – namely part-of-speech (POS) tagging and grammatical dependency parsing. The dependency parsing is most informative on

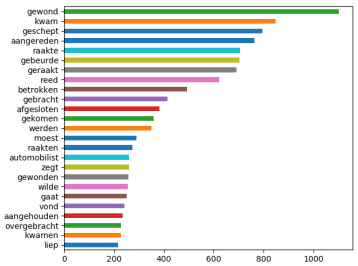
isolated sentences and not on the data set as a whole. POS tagging does give insight in key terms in the data and the importance of VRU and objects in the data. POS tagging was done with spacy and trained on the Dutch alpino training set.[8] The POS tags in the title and text features were plotted, most frequent adjectives; nouns; and verbs are shown in figure 4 and 5. As can be seen in both the figures the VRU and objects are some of the most frequent nouns in the data. However, there are some inaccuracies in the POS tagging. In the text, 'rotterdam' and 'breda' are incorrectly categorised as adjectives. In the titles, there are some more inaccuracies: again 'breda' is categorised as an adjective; 'motorrijder', 'fietser', 'scooterrijder' and 'paalbericht' are also incorrectly categorised as adjectives. Furthermore, 'rotonde', 'utrecht' and 'breda' are incorrectly categorised as verbs. The difference in accuracy between the text and title is most likely because titles are not always correct sentences which makes it hard to correctly annotate the terms.



(a) Most frequent adjectives



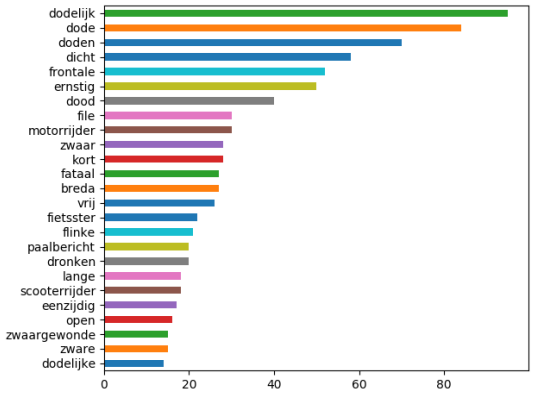
(b) Most frequent nouns



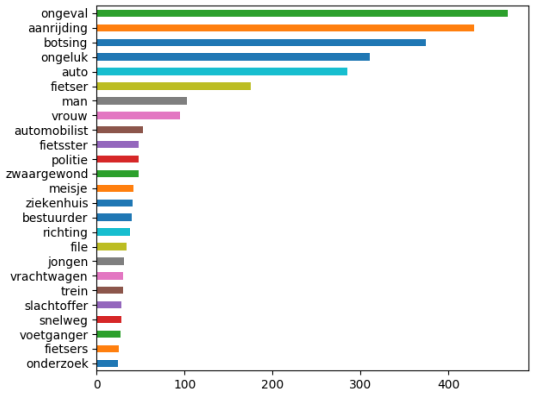
(c) Most frequent verbs

Figure 4: Most frequent nouns, adjectives and verbs in text

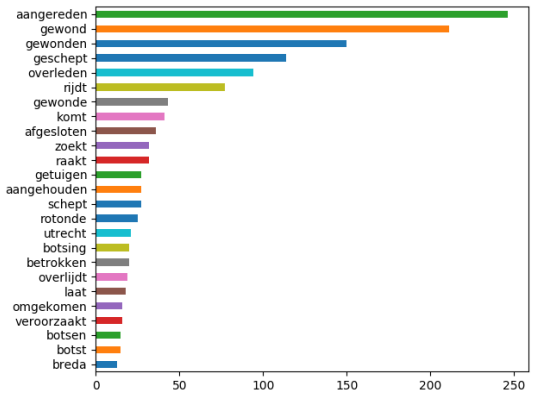
[8] https://spacy.io/



(a) Most frequent adjectives



(b) Most frequent nouns



(c) Most frequent verbs

Figure 5: Most frequent nouns, adjectives and verbs in titles

4.3 Method: potential pattern finding

The goal of the pattern analysis was to prove that there are patterns in the data and that these can be captured in computational terms. The following steps were taken to identify a pattern:

1. Manually look through the data and find a potential pattern to analyse
2. Analyse this pattern in terms of POS tags and grammatical dependencies (POS tags individually are not very informative and Noun Chuck parsing might have been beneficial as will be mentioned in the discussion section)
3. Define the pattern in the most promising pattern i.e. POS tags or grammatical dependencies
4. Use different levels of abstraction to gauge the validity of the pattern

The pattern meant above are relations between specific entities, which are based on the previously mentioned research in discursive content analysis. The target relations between these entities consist of relations as for example "X hits Y". The term 'hits' can be many different terms in dutch as for instance 'raakt', 'schept', or 'aangereden'.

There are two categories of entities between which relationships are detected, vulnerable road users (VRU) and objects. The category of VRU was defined by the following collection of terms :'fietser', 'man', 'vrouw', 'fietsster', 'voetganger'. In table 2 the occurrences of the VRU are shown in respect to titles and text. In the text feature the corpus frequency was taken, which is why these terms were more frequent. The category of objects in this research only consist

	Title occurrences	Text occurrences
Fietser	175	548
Fietsster	70	157
Man	103	1169
Vrouw	95	870
Voetganger	40	93

Table 2: Occurrences of VRU in title and text

of the term: 'auto'. This is because this is the most prominent object in news articles and which is involved in almost all reported traffic accidents. As can be seen in table 3, car is a frequent term in our data and used more than any of the VRU. For comparison, the table also shows the occurrence of the term 'automobilist' which is the more personal term of car but used far less.

	Title occurences	Text occurences
Auto	288	2357
Automobilist	70	626

Table 3: Occurrences of 'auto' vs 'automobilist' in titles and text

5 Results

In this section, one pattern has been worked out and statistically substantiated.

1. Following the defined method, first, a potential pattern was manually identified. Looking through the titles of sentences, there were frequent sentences similar to form of: 'fietser geschept door auto'. In total, there were 51 titles which contained the term 'auto' and 'geschept'.

2. The second step is analysing the titles in term of the POS tags and grammatical dependencies. In the following case, only the grammatical dependencies were shown to be significant and POS tags did not show a pattern. Three main dependencies were found to be of significance in the sentences: nominal subjects, oblique nominals and root verbs. The meaning of these dependencies is as follows: the root is the core relation in a sentence; the nominal subject is the the typical subject of a sentence;the oblique nominal is a noun identified as not the core noun.[9] For the 51 titles, 48 of them contained as 'geschept' as the root verb. The nominal subject of a sentence corresponded to the victim of an accident 46 times out of 51. In addition, 33 of these times the victim was a VRU. The most common nominal subjects can be seen in figure 6. The oblique nominal corresponded to the term 'auto' 43 times, in all these cases the car was involved in the accident. The most common oblique nominals can be seen in figure 7. There are some cases where the oblique nominal did not correspond to someone/something involved like the term 'ziekenhuis'. Furthermore, 32 titles contained multiple oblique nominals with one of them always being 'auto', the other ones contained term not involved with the accident like the previously mentioned 'ziekenhuis'.

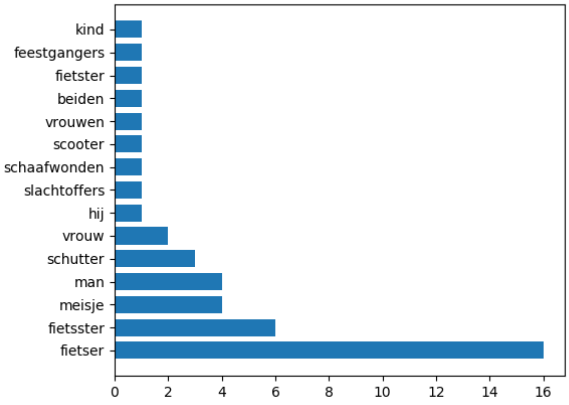


Figure 6: Most common nominal subjects in 'geschept' and 'auto' titles

[9]Further explanations can be found here: `https://universaldependencies.org`

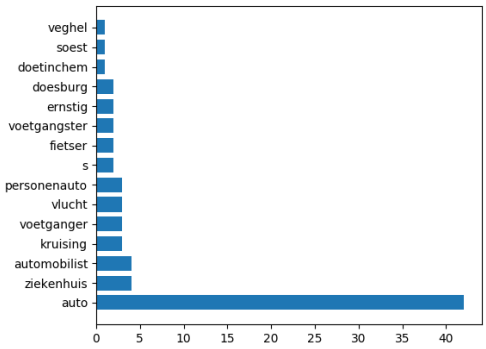


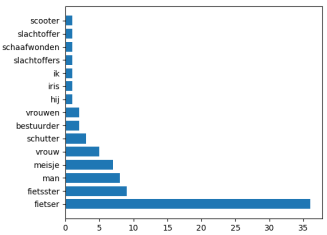
Figure 7: Most common oblique nominals in 'geschept' and 'auto' titles

3. The third step is to define a potential pattern, in this case in grammatical dependency terms. The grammatical dependency pattern is defined as:

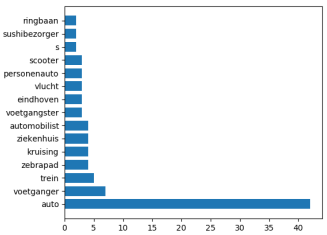
nominal subject * root * oblique nominal

This means that a sentence would start with a nominal subject possibly some terms in between, then the root verb and the oblique nominal. Intuitively, returning to the first step, we are looking for sentences as for example "Fietser geschept door auto" or "Voetganger aangereden door auto". The assumption is that the nominal subject and oblique nominal are persons/objects involved in the accident and the root the relation as to how they are involved.

4. The last step is to see how the potential pattern performances on different cases. Firstly, titles with only the term 'geschept' were collected. In total, 117 titles contained the term with 98 having 'geschept' as the root grammatical dependency. The most common nominal subjects and oblique nominals for these 98 sentences are shown in figure 8. The nominal subjects are similar to



(a) Most common nominal subjects



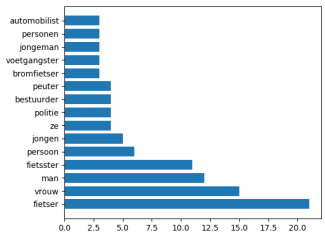
(b) Most common oblique nominals

Figure 8: Nominal subjects and oblique nominals in 'geschept' sentences

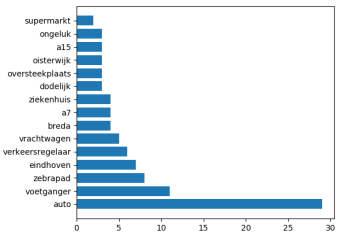
figure 6, with nominal subjects corresponding to entities which were involved in the accident expect for the subject 'schaafwonden'. The oblique nominals are also similar to the results of the previous step, as can be seen in figure 7. They

contain the same faults with terms like ‘ziekenhuis’, ‘kruising’, ‘zebrapad’ which are not entities involved in the accident.

Secondly, all titles with the term ’aangereden’ were collected. In total, there were 226 titles with the term and of those 188 had ’aangereden’ as the root grammatical dependency. The most common nominal subjects and oblique nominals for these 188 titles are shown in figure 9. For the term ‘aangereden’ there are still similar results to those with the term ‘geschept’.



(a) Most common nominal subjects



(b) Most common oblique nominals

Figure 9: Nominal subjects and oblique nominals terms in ’aangereden’ sentences

Lastly, all titles were analysed for their nominal subjects, oblique nominals and roots. The two most common roots are ‘aangereden’ and ‘geschept’ which were already discussed. Three other common roots were ‘gewond’, ‘rijdt’ and ‘overleden’ of which the most common nominal subjects and oblique nominals are shown in tables 4,5 and 6. As seen in the tables, nominal subjects correspond to entities involved in the accident in all cases. However, the oblique nominals do not correspond to entities involved in the accident and in most cases actually refer to a term describing the accident as for example ‘aanrijding’ and ‘botsing’. This shows that the defined pattern is not an applicable pattern to all titles and is only relevant for some specific cases like ‘geschept’ and ‘aangereden’.

Nominal subjects	Frequency
Fietsster	18
Vrouw	13
Motorrijder	10
Automobilist	8
Fietser	7

Oblique nominals	Frequency
Aanrijding	33
Botsing	30
Ongeluk	10
Ongeval	8
Ernstig	6

Table 4: Most common nominal subjects and oblique nominals in relation to ‘gewond’

Nominal subjects	Frequency
Man	13
Meisje	8
Vrouw	8
Fietser	6
Fietsers	5

Oblique nominals	Frequency
Aanrijding	19
Ongeval	14
Botsing	10
Ongeluk	9
Enchede	6

Table 5: Most common nominal subjects and oblique nominals in relation to 'overleden'

Nominal subjects	Frequency
Automobilist	14
Auto	10
Man	9
Bestuurder	5
Luxemberger	1

Oblique nominals	Frequency
Aanrijding	21
Botsing	12
Ongeval	9
Fietser	8
Pilaar	6

Table 6: Most common nominal subjects and oblique nominals in relation to 'rijdt'

6 Conclusion

Returning to the central question of this thesis: how can computational linguistics techniques be used to find patterns between entities within traffic accident news? As shown in computational approaches section, there has been research into computational linguistics techniques which solve similar problems. The unsupervised learning algorithms like that of Chambers and Jurafsky (2011) find patterns between entities without and predefined structure, also refered to as template induction. Therefore, it is possible to automate content analysis, finding patterns between entities, using these techniques. To use these techniques on the Dutch traffic accident domain the resources needed for these techniques have to be compatible. This means that a Dutch NER would have to be used and a Dutch wordnet version configured. Using this unsupervised approach could be beneficial to discursive content analysis, making it possible to automatically identify relations and patterns on a large scale.

The possibility of using these technique in the traffic accident domain depends on viable patterns being present in the domain. The exploratory pattern analysis showed that there are some patterns to be found in the data. In the results section the example of the term 'geschept' was taken and it was shown that looking at the computationally annotated grammatical dependencies a pattern does emerge. In this pattern, the nominal subject and oblique nominal consistently corresponded to those involved in the accident. Furthermore, the nominal subject was also found to be the victim – in the sense of the one being hit – in most cases. Tested on all titles, the pattern did not emerge in all examples showing that only for some terms are relevant for this pattern. This is expected since there is not just one pattern in the text but multiple where some are more significant then others. The discursive content analysis described in the theoretical context aimed to identify similar relations, finding who had been hit and

which terms were used to describe them. This shows that interesting discursive patterns can be found in the traffic accident domain, potentially through computational means like unsupervised template induction.

7 Discussion

In terms of the method and results there are a number of caveats. The collected data had several faults. Lexis Nexis has data on far more then only the domain which this research is interested in, which means that when searching for articles there is no assurance that the article is actually relevant. The search queries used for the selecting of articles are relevant but there are too many articles to manually calculate precision. Therefore, the data could consist of false positives which skews the results. Secondly, the journal annotations contained some variations in the journal name, as for example "Copyright 2012 de persgroep" and "Copyright 2013 de persgroep". The actual journal in this case is "de persgroep", these variations made the distribution of journal annotations less accurate.

The found pattern in the results section was only analysed through the titles. The reason for this is that the sentences in the main text of an article tend to be a lot more complicated then the titles. Simple patterns and grammatical dependency structures are therefore hard to find in the main text. This can also be seen in the examples given by Ralph et al. (2019) in their research, where complicated sentences were categorised in simpler terms. On then contrary, titles only consist of about 5 words with the main cause and people involved in an accident frequently reported. However, titles are often not whole sentences which makes it harder to find the correct grammatical annotations. A promising approach is to use the main text and employ noun chunk parsing to find parts of sentences which are of importance. Due to the complexity and variety of sentences in the articles this was not done in this research.

In future work it would be interesting to actually transfer the research done by Chambers and Jurafsky (2011) in a Dutch traffic accident domain. Furthermore, specifically recreating the research done by (Ralph et al., 2019) in a Dutch content could be informative. Focusing less on the automation through machine learning techniques – as was done in this research – of discursive content analysis but instead using techniques like noun chuck parsing.

References

Aarts, L., Eenink, R., Weijermans, W., Knapper, A., & Schagen, I. (2014). *Soms moet er iets gebeuren voor er iets gebeurt* (Tech. Rep.). Stichting Wetenschappelijk Onderzoek Verkeersveiligheid {SWOV}. Retrieved from `https://www.swov.nl/en/publication/soms-moet-er-iets-gebeuren-voor-er-iets-gebeurt`

Agichtein, E., & Gravano, L. (2000). Snowball: Extracting relations from large plain-text collections. In *Proceedings of the fifth acm conference on digital libraries* (pp. 85–94).

Allahyari, M., Pouriyeh, S. A., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. J. (2017). A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. In *Proceedings of kdd bigdas* (p. 13).

Chambers, N., & Jurafsky, D. (2011). Template-based information extraction without the templates. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1* (pp. 976–986).

Daniels, S., Brijs, T., & Keunen, D. (2010). Official reporting and newspaper coverage of road crashes: A case study. *Safety Science*, *48*(10), 1469–1476. Retrieved from `http://linkinghub.elsevier.com/retrieve/pii/S0925753510001931` doi: 10.1016/j.ssci.2010.07.007

De Ceunynck, T., De Smedt, J., Daniels, S., Wouters, R., & Baets, M. (2015). "Crashing the gates" - Selection criteria for television news reporting of traffic crashes. *Accident Analysis and Prevention*, *80*, 142–152. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0001457515001396` doi: 10.1016/j.aap.2015.04.010

Houwing, S. (2017). *De beschikbaarheid en kwaliteit van informatie over verkeersongevallen* (Tech. Rep.). Instituut Wetenschappelijk Onderzoek Verkeersveiligheid. Retrieved from `https://www.swov.nl/publicatie/de-beschikbaarheid-en-kwaliteit-van-informatie-over-verkeersongevallen`

Huang, R., & Riloff, E. (2012). Bootstrapped training of event extraction classifiers. In *Proceedings of the 13th conference of the european chapter of the association for computational linguistics* (pp. 286–295).

Jiang, J. (2012). Chapter 02 - Information extraction from text. In *Mining text data* (pp. 11–41). Springer. Retrieved from `http://www.springerlink.com/index/10.1007/978-1-4614-3223-4` doi: 10.1007/978-1-4614-3223-4

Kemler, H. J., & den Hertog, P. (2009). Ongevallen in Nederland: Overzicht en risicogroepen per ongevalscategorie. *Tijdschrift voor gezondheidswetenschappen*, *87*(8), 360–365. Retrieved from `http://link.springer.com/10.1007/BF03082302` doi: 10.1007/BF03082302

Liddy, E. D. (2001). Natural Language Processing. In *Encyclopedia of library and information science* (2nd ed.). NY: Marcel Decker, Inc.

MacRitchie, V., & Seedat, M. (2008). Headlines and discourses in newspaper reports on traffic accidents. *South African Journal of Psychology*, *38*(2), 337–354.

Magusin, H. (2017). If You Want to Get Away with Murder, Use Your Car: A Discursive Content Analysis of Pedestrian Traffic Fatalities in News

Headlines. *Earth Common Journal*, *7*(1), 28–62.

Mejia, P. (2017). *Beyond the traffic report: The news about road safety and vision zero in San Francisco* (Tech. Rep.). Berkeley Media Studies Group. Retrieved from http://www.bmsg.org/sites/default/files/bmsg_vision_zero_traffic_safety_report_2017-01-09.pdf

Ralph, K. M., Lacobucci, E., Thigpen, C., & Goddard, T. (2019). Editorial Patterns in Bicyclist and Pedestrian Crash Reporting. In *Presented at the transportation research board 97th annual meeting, wshington dc.* Retrieved from TRBPaperNO.19-03892http://www.eden.rutgers.edu/~ei60/crashespaper.pdf

Rijkswaterstaat. (2018). *Handleiding product Bestand geRegistreerde Ongevallen Nederland (BRON)* (Tech. Rep.). Rijkswaterstaat. Retrieved from https://www.rijkswaterstaat.nl/apps/geoservices/geodata/dmc/bron/Documentatie/HandleidingproductBestandgeRegistreerdeOngevallenNederland.pdf

Rosenfeld, B., & Feldman, R. (2007). Clustering for unsupervised relation identification. In *Proceedings of the sixteenth acm conference on conference on information and knowledge management* (pp. 411–418).

Scheufele, D. A., & Tewksbury, D. (2006). Framing, agenda setting, and priming: The evolution of three media effects models. *Journal of communication*, *57*(1), 9–20.

Shinyama, Y., & Sekine, S. (2006). Preemptive information extraction using unrestricted relation discovery. In *Proceedings of the main conference on human language technology conference of the north american chapter of the association of computational linguistics* (pp. 304–311).

Sorokin, D., & Gurevych, I. (2017). Context-Aware Representations for Knowledge Base Relation Extraction. In *Proceedings of the 2017 conference on empirical methods in natural language processing (emnlp)* (pp. 1784–1789). Association for Computational Linguistics. doi: 10.18653/v1/D17-1188

Van den Bosch, A., Busser, G. J., Daelemans, W., & Canisius, S. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch. (Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting), 99–114.

Verkade, T. (2017). Door deze fietsprofessor kijk je voor altijd anders tegen het fileprobleem aan. Retrieved from https://decorrespondent.nl/7116/door-deze-fietsprofessor-kijk-je-voor-altijd-anders-tegen-het-fileprobleem-aan/748226585700-3d826dc2

Verkade, T., & Brömmelstroet, M. t. (2018). Honderden doden, duizenden gewonden, tienduizenden trauma's: verkeersleed is akelig dichtbij. Retrieved from https://decorrespondent.nl/8757/honderden-doden-duizenden-gewonden-tienduizenden-traumas-verkeersleed-is-akelig-dichtbij/920772935775-6bf83c40

Xie, Q. (2018). Critical discourse analysis of News Discourse. *Theory and Practice in Language Studies*, *8*(4), 399–403. doi: 10.17507/tpls.0804.06