“What a Thesis!”: A Study on the Characterization and Detection of Exclamatives in Sentiment Analysis

Student: Dignée Brandt
Begeleider: dr. Rick Nouwen
Tweede Beoordelaar: prof. dr. Yoad Winter
Datum: 25.06.2016

Bachelor Artificial Intelligence, UU
7,5 ECTs

Table of Contents

Introduction

When someone says, “my new phone broke after three hours”, he or she does not explicitly express a judgement, but nevertheless expresses an opinion that is similar to “this phone’s quality is horrible!”. The hearer intuitively understands this implicit meaning, will be disappointed and express his or her compassion with the hearer, and will keep in mind to never buy that particular phone.

How does the hearer understand the speaker’s opinion, while the latter does not explicitly express it? This is one of the questions that occupies linguistics. More specifically speaking, the question of *how* the speaker expresses such an opinion, is one for the field of sentiment analysis. The same goes for more explicit exclamatives, such as “this is amazing!” or “I love this cake!”. What kind of sentiment do these exclamatives denote, and how can we explicitly categorize them? Is there a fundamental linguistic difference between the expression of negative opinions and positive opinions? These are all question that are involved when researching sentiment analysis.

In this thesis, I will try to find an answer to the following question: what role do exclamatives play in the context of sentiment analysis, and is there an algorithmic way to detect exclamatives and categorize them? This question has both linguistic aspects and applications in artificial intelligence: natural language processing is one of the main topics that the occupies research in the latter area.

I will try to find an answer to this question in the following way. In chapter 1, I will go over the general concepts of sentiment analysis and exclamatives. What do these concepts mean, and how do they relate to each other? In chapter 2, I will discuss an experimental method about exclamatives that characterizes them in terms of statistical models. After a thorough discussion of this method and more general challenges in corpus-related experiments, I will perform a similar version of the experiment on a Dutch corpus in chapters 3 and 4. This will all lead to a conclusion that argues for promising possibilities regarding the automatizing of (parts of) sentiment analysis.

Chapter 1: Sentiment Analysis and Exclamatives

When writing a thesis in the field of sentiment analysis, the first thing to address is the concept of sentiment analysis itself. After clarifying what this concept means, I will discuss what the goals of sentiment analysis are, go over some of the related main notions, the obstacles of sentiment analysis and the importance and relevance of it in the broad field of linguistics. After this, I will shift my focus to the main aspect of sentiment analysis I will be discussing in this thesis: exclamativity. I will argue how we should understand it in this context, and why it is an important aspect in sentiment analysis.

1.1 Sentiment Analysis

What is sentiment analysis?

In the fields of linguistics and natural language processing, sentiment analysis is the study of an extra meaningful layer of language. This layer consists of, but is not limited to, emotions, opinions, judgements, and the speaker's or writer's attitude. It is a rather large subject space, but nevertheless an important aspect of language.[1] The goal of sentiment analysis is to understand the different ways one could use to express all these sorts of sentiments when using language. One advantageous aspect of sentiment analysis is that it limits itself to the analysis of the emotional or sentimental layer of language, and thus does not have to understand a complete system of semantics.[2] There is no need to understand the complete meaning of a certain phrase, the only necessary information to extract is the emotional meaning.

A more specific goal of sentiment analysis is to not only understand how sentiment could be expressed through language, but exactly *what* kind of sentiment is expressed. The distinction to be made is generally binary: either the expressed sentiment is positive, or the expressed sentiment is negative.

For a clearer understanding and analysis of some of the aspects of sentiment analysis, it is important to define some of the concepts that I will use throughout this thesis. Another term for sentiment analysis is 'opinion mining', although the former is more commonly used than the latter. There are some differences between the two, but these differences are not relevant in this thesis. This is because in this thesis, I take the spectrum of sentiment analysis to be very large, thus including opinion mining. For the sake of clarity, I will consistently use the term sentiment analysis, also when referring to literature that uses the term 'opinion mining'.[3]

A second important term is 'polarity'. As mentioned, sentiment analysis pursues the distinction between positive expressions and negative expression.[4] This is precisely what the term polarity holds: when speaking about the polarity of a phrase, one speaks about the aspect of the phrase that expresses either a positive or a negative emotion.

---

[1] Bing Liu, *Sentiment Analysis and Opinion Mining,* Morgan & Claypool Publishers (2012), 7.
[2] Ibid., 13.
[3] Bo Pang and Lillian Lee, "Opinion Mining and Sentiment Analysis," *Foundations and Trends in Information Retrieval* 2/1-2 (2008), 1-135, https://www.cse.iitb.ac.in/~pb/cs626-449-2009/prev-years-other-things-nlp/sentiment-analysis-opinion-mining-pang-lee-omsa-published.pdf (retrieved June 10th, 2016), 9-10.
[4] Ibid., 10.

A third important concept is that of 'sentiment'. When using this term, I mean all the emotional attitudes, judgements and opinions that can be attached to a certain linguistic expression.

While engaging in sentiment analysis, it is important to note on which linguistic level one is analysing: for each level, there are certain goals and difficulties. In his 2012 book *Sentiment Analysis and Opinion Mining,* Bing Liu distinguished three levels of sentiment analysis:[5]

(1) The document level: the goal of research on this level is to determine whether the *whole piece* of text you are analysing is either positive, negative or neutral. The assumption here is that this piece of text expresses only one kind of sentiment, and is thus "not applicable to documents which evaluate or compare multiple entities."[6]

(2) The sentence level: the goal of research on this level is to determine whether a sentence is either positive, negative or neutral. As well as for the document level, the assumption is that a sentence can convey only one judgement (positive, negative or none at all).

(3) The entity or aspect level. Analysis of this level results in "finer-grained analysis".[7] At this level, it *is* possible to express more than one judgement in a piece of text. Analysis here will distinguish between *judgements* and *subjects of judgements*, resulting in a nuanced and specific analysis of expressed opinions. Analysis on this level is the most useful, as it can provide more specific results out of more complex pieces of text (i.e. text that express multiple emotions and opinions), but is at the same time the most difficult *because of* the complexity.

This thesis mainly focuses on the entity or aspect level, but relates it to the other two levels, by showing how the former could be useful for analysis of the latter. Because of the complexity of the entity/aspect level, I will now discuss some of the main problems that arise when engaging in sentiment analysis. The problems all arise from the versatility of language and the countless ways of expressing yourself: of course an emotional expression can be expressed by simply saying "I hate x", but there are also a lot of rhetorical ways, figures of speech and other uses of language that achieve the same result. For example, someone could say, "what a restaurant!", a sentence that does not *explicitly* convey emotion, but nevertheless clearly express an extreme opinion (either the speaker really likes or really hates the restaurant).

Difficulties regarding sentiment analysis

In this overview, I want to discuss the (in my opinion) four most important problems that arise in sentiment analysis:

(1) A recurring fact that is problematic, is that there is no such thing as a fixed syntactic model for sentences that express sentiment. Although there are words that are

[5] Liu, *Sentiment Analysis and Opinion Mining,* 10-12.
[6] Ibid., 11.
[7] Ibid., 11.

presumably often used when expressing sentiment, the use of them is not a general rule. Thus, there are "many sentences without sentiment words" that "can also imply opinions,"[8] like the mentioned example of "my new phone broke after three hours." This makes it harder to distinguish sentences that do express sentiment from the sentences that do not. For example, a sentence like "the screen of my phone cracked after one day" is a seemingly objective, factual statement, but obviously expresses a negative opinion.

(2) The problem in (1) also applies the other way around: "a sentence containing sentiment words may not express any sentiment."[9] For example, the sentence "I am looking for a good and funny movie" does contain the seemingly positively loaded words 'good' and 'funny', but the sentence as a whole does not express an opinion or emotion. This again enlarges the set of potential sentences that express sentiment.

(3) A third problem for sentiment analysis is sarcasm. Sentences that seem to be perfectly clear expressions of an opinion, in that case convey the exact opposite emotion. Let us take a sentence like "wow, I am so blessed to receive over a hundred angry phone calls a day, I am lucky to have this amazing job at the customer services!" seems to express a positive emotion ('blessed', 'lucky', 'amazing') but obviously entails a rather negative one. It is still not clear what exact aspect of these kind of sentences makes them so obviously sarcastic, and thus very hard to automatically characterize: the mentioned sentence does not have any explicit signs of negativity, but should be categorized that way.

(4) A last problem for sentiment analysis is the fact that people express their sentiment in different ways on different platforms.[10] When writing an official complaint to a company about a product, people tend to express the same emotion in another way than when telling about it to your friends in a bar. This makes it hard to characterize specific ways of expressing sentiment, as there does not seem to be one general pattern of expression.

I will address some of these difficulties in both the discussion of the Potts and Schwarz method and in the discussion of the bol.com experiment, especially points (1) and (4). The other points are in my opinion equally interesting, but fall outside of the scope of this thesis. However, these still are interesting and important points to consider when studying sentiment analysis, something I wish to recommend to focus on in future studies on the topic.

Relevance and importance

Before shifting my focus to exclamativity, the main topic of this thesis, I want to situate sentiment analysis in the broader field of linguistics. Why is research in sentiment analysis relevant and important?

The answer is twofold: there is both relevance in the field of theoretical linguistics as well as in the pragmatic sense, especially in commercial applications and in the field of

[8] Liu, *Sentiment Analysis and Opinion Mining,* 13.
[9] Ibid., 12.
[10] Ibid., 16-17.

artificial intelligence. As for the former, research in sentiment analysis provides rich information about this specific aspect of semantics. It could potentially enrich our understanding of natural languages and the variety of linguistic ways of expressing particular emotions. Some of the results of this search for understanding is already visible, for example in projects as SentiWordNet.[11] This is a lexical database consisting of the earlier existing lexical database WordNet, which annotates words on meaning and similarly offers synonyms. SentiWordNet adds an annotation of sentiment, divided over three categories: positive, negative and objective. These three categories each have a score between (including) 0.0 and 1.0, the sum of all three scores is 1.0. Each score indicates the extent of association of that word with positivity, negativity or objectivity. This gives a solid indication of the commonly associated sentiment with that word (or the absence of it). SentiWordNet seems to be rather valuable and accurate, as it is used by "more than 300 research groups and used in a variety of research projects worldwide."[12]

However, the relevance and importance for commercial and AI purposes is way more extensive. In fact, the rise of technology in the late twentieth and early twenty-first century has been an important factor for the expansion of the field of sentiment analysis.[13] This is partly because of the need for automation: it initiated a huge demand for natural language processing systems, where sentiment analysis obviously is an important part of. Commercial companies were also eager to analyse their clients' opinions which were widely expressed on the internet. Another cause of the expansion of the field is because through this technological rise, in particular the ascent of the world wide web, there were suddenly large sets of data available and ready to use. This stimulated the research on machine learning applications, as well as some part of the field of robotics.

Nowadays, these are still the main categories in which sentiment analysis is being used. Apart from the diverse commercial applications, state of the art artificial intelligence research and large companies as Google and Amazon enhance their intelligent systems while among other things using results of sentiment analysis research. Applications like voice assistance on mobile phones benefit from this area of research, as well as analysis of big data – either for governmental and commercial purposes. A big challenge for artificial intelligence is nowadays still the understanding of non-explicit information, like emotions and attitudes. Sentiment analysis will largely contribute to the linguistic side of the possible solutions, as the implicit will be turned into the explicit with the help of smart algorithms that have a sentiment analysis-drive basis.

1.2 Exclamatives and Exclamativity

I will now turn to the main topic of the rest of the thesis, namely the topic of exclamativity. My goal is to study exclamatives, because, as I will argue below, they are *the* subject of sentiment analysis: thus, the first step into the study of sentiment analysis should in my opinion be the study of exclamatives. I will do this by first addressing the following

[11] Stefano Baccianella et al., "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," *Proceedings of the Seventh Conference on International Language Resources and Evaluation* (LREC'10) (2010), 2200–2204.
[12] Baccianella et al., "SentiWordNet 3.0," 2200.
[13] Pang and Lee, "Opinion Mining and Sentiment Analysis," 7-8.

questions: what are exclamatives exactly, how do they fit into the field of sentiment analysis, and why are they of importance? After this, I will discuss some of the relevant aspects of exclamatives, and introduce the second part of the thesis: Christopher Potts and Florian Schwarz's experimental method of studying exclamatives. I will study exclamativity according to this method by, amongst other things, applying it to my own Dutch corpus.

What are exclamatives exactly? Exclamatives are in fact the exact limited subject of sentiment analysis: they are those items, words, phrases or sentences, that convey *some* kind of sentiment. Exclamatives are those items that are not merely declarative or neutral expressions of language.[14] It is important to highlight the distinction between *an exclamative* and *exclamativity*: the former is the bearer of the sentiment, in most cases a morpheme, word or phrase. The latter is the sentiment that is being conveyed when expressing an exclamative.

The question of how we should understand exclamatives in the context of sentiment analysis is thus a relatively easy one. Exclamativity is precisely that which the sentiment analysis is looking for – the exclamatives are the ultimate subject of sentiment analysis (ultimate meaning the relevant set of subjects that remains after distinguishing between items that convey sentiment and items that do not). It is thus safe to say that exclamatives play a large and important role in the field of sentiment analysis.

However, there are other aspects of exclamatives that need to be highlighted here. It is important to realise that the detecting of exclamatives is the crucial first step of sentiment analysis. Therefore, it is of great importance that there is an accurate way of classifying words and phrases as exclamatives. This minimizes the dataset that sentiment analysis has to deal with.

Because of the fact that the detecting of exclamatives is the first step of sentiment analysis, there are less problems involved that have to be dealt with. There is not yet a question of polarity involved, neither is the problem of sarcasm of relevance. However, this is not to say that the problems of sentiment analysis are hereby resolved. Exclamatives can only express so much, not including detailed results about the polarity of linguistic items.

This fact seemed to motivate Christopher Potts and Florian Schwarz to do research on exclamatives.[15] Focusing on the role exclamatives play in sentiment analysis, the next chapter will discuss their experimental method to detect and characterize exclamatives in depth. Hereafter, I will discuss in detail the experiment they performed on an English corpus of reviews and perform it myself on a Dutch corpus. In the end, I hope to achieve a better understanding of the use of exclamatives throughout different languages, and acknowledge their importance for research in sentiment analysis.

[14] Anna Chernilovskaya, *Exclamativity in discourse: Exploring the exclamative speech act from a discourse perpective,* Dissertation Utrecht University (2014), 2.
[15] Cristopher Potts and Florian Schwarz, "Exclamatives and heightened emotion: Extracting pragmatic generalizations from large corpora," Ms., UMass Amherst, 2008.

Chapter 2: Potts and Schwarz, Towards Characterizing Exclamatives

Following up on the theoretical background of sentiment analysis and the included exclamativity discussed in chapter 1, I will now turn to a particular method for studying exclamativity, due to Christopher Potts and Florian Schwarz (2008).[16] This experimental method takes a closer look at exclamativity and its role in the context of sentiment analysis. I will first address the goal of the method and the nature of exclamativity. Then I will address the conducted experiment in detail, and discuss how this is relevant in the bigger picture of exclamativity and sentiment analysis.

2.1 Goals and Context

The goal of the Potts and Schwarz method is to investigate "how to characterize"[17] a particular segment of sentiment analysis, namely that of exclamativity. They want to broaden the understanding of the concept of exclamativity, and find characterizing and unique aspects of it. An exclamative is, in this pragmatic context (pragmatic being focused on the use-conditions of language), some meaningful piece of language with an extra emotional charge ("encodes excitement, surprise," et cetera).[18] A more detailed definition can be found in chapter 1.

Potts and Schwarz begin by naming a first important aspect of exclamatives, namely that there are generally two linguistic ways to express exclamativity. The first way is to express it by using "certain adverbials [that] can also layer an exclamative semantics atop a declarative foundation."[19] This means that one could take a neutral adverbial (phrase) and then use it in a certain way or context to give it an extra semantic, emotional layer. The example that Potts and Schwarz use is the phrase 'what a'. This phrase is on itself neutral, in a sentence like "this is what a typical 18th century painting looks like." However, the neutral phrase could also be used as an exclamative, in a sentence like "what a hotel!" In this sense, 'what a' conveys an emotional charge, namely that the speaker finds the hotel either really good or really bad.

The second way of expressing exclamativity is by using a "variety of particles whose sole function is to convey [...] pure exclamativity"[20] This means that one could use a word or phrase that cannot be used neutrally, as we saw at the previous example of 'what a', but is per definition emotionally charged. Examples of these words or phrases are 'wow' or 'oh my god'.

An important property of exclamativity is that the only vast characterization that can be made of it, is that it conveys *some* heightened emotion. There is "no single meaning associated with exclamatives" but "generalized heightened emotion."[21] In practice, this means that even though we might be able to characterize aspects of exclamatives, we cannot immediately

[16] Potts and Schwarz, "Exclamatives and heightened emotion."
[17] Ibid., 3-4.
[18] Ibid., 3.
[19] Ibid., 4.
[20] Ibid.
[21] Ibid., 5.

know if these exclamatives are positively or negatively charged. An exclamative usually does not have a fixed polarity, meaning that it usually does not express positivity per se or negativity per se: most of the time it can express both, and the polarity can only be determined by studying the context.

This is not to say that it is impossible to make some distinction in the heightened emotions. Potts and Schwarz propose a way to roughly distinguish between three types of exclamativity: positive, negative and both (and, of course, no exclamativity at all). This results into four different statistical profiles for exclamatives: positively biased, negatively biased, and 'general' (both positive and negative) exclamatives; with the fourth profile denoting the absence of exclamativity.

After these observations on exclamatives, Potts and Schwarz introduce their experiment that will help them characterize exclamativity. They perform a statistical analysis on a large corpus, consisting out of online product reviews, each accompanied with a rating scale from one to five stars. To put it very briefly, they propose characterizing statistical profiles for exclamativity, by relating particular rating-scores to particular linguistic expressions in the reviews.

I will first discuss the hypotheses, then the corpus, and explain the statistical analysis that they use on the data in the corpus. After that, I will discuss the results and shortcomings, as well as a brief overview on how this experiment helped them in their goal to characterize exclamativity.

Hypotheses

Potts and Schwarz first propose some intuitive hypotheses about exclamatives:[22]

> (h1) "Speakers writing one-star or five-star reviews are (or seek to create the impression that they are) in more heightened emotional states than speakers who are writing two, three or four star reviews."
> (h2) "A speaker who uses an exclamative is in a heightened emotional state (or at least seeks to create such an impression)."

Together, these two hypotheses form their main hypothesis:

> (h3) "Exclamatives are more frequent in reviews with extreme ratings (both positive and negative)."[23]

Potts and Schwarz take this hypothesis (h3) as a starting point towards identifying "a general statistical profile for exclamatives":[24] assuming that h3 is true, they try to retrieve this in the corpus, in the form of specific statistical profiles that indicate the presence of exclamativity. "This [the statistical profiles] allows us to locate all the exclamatives in a corpus with this

[22] Potts and Schwarz, "Exclamatives and heightened emotion," 16.
[23] Ibid.
[24] Ibid., 17.

structure, without appeal to native speakers' intuitions, a close reading of the text, or deep understanding of the context."[25] The definitive goal of their experiment is thus to find statistical profiles of words or phrases that imply that these words or phrases are exclamatives.

2.2 The Corpus and the Experiment

The Potts and Schwarz corpus

The Potts and Schwarz corpus exists out of approximately 100.000 English reviews of hotels and books from the websites TripAdvisor.com and Amazon.com, with an accompanying rating scale of 1 (worst) to 5 (best).[26,27] The relevant parts of the reviews that have been put into the corpus are (1) a one-sentence summary of the review, written by the author of the review; (2) the review itself; (3) the rating. This adds up to four separate corpora: the TripAdvisor.com summaries, the TripAdvisor.com reviews, the Amazon.com summaries and the Amazon.com reviews.

The data in each corpus has been systematically categorized by both unigram-, bigram- and trigram-tokens. For each token with length *i* in each rating-category ('ever' in 1-star reviews, 'ever' in 2-star reviews, et cetera), there are two values attached: (1) the amount of occurrences of this token in this rating category, the *tokencount*; (2) the absolute amount of length-*i* tokens in this rating category over all, the *widecount*. A line of data in the corpus then looks as follows:

Word	*Rating*	*Tokencount*	*Widecount*
'ever'	1	371	570687

This line means that the token 'ever' occurs 371 times in 1-star reviews, with a total of 570687 unigrams in all 1-star reviews.

Statistics

The underlying idea for the experiment is rather simple: Potts and Schwarz take a look at the frequency of specific tokens in reviews and filter out relevant frequency-patterns. This could for example be a pattern of a token that has a lot of occurrences in 1- and 5-star reviews, and less occurrences in 2-, 3- and 4-star reviews. According to hypothesis (h3), this specific token should convey a high level of emotion.

However, to really say something reliable about these tokens, the statistics need to be a little more complex. When just looking at frequencies, the image could get distorted, as the corpus is not a perfectly balanced representation of the use of language. It could be that the involved reviews are coincidentally rather negative, or maybe it coincidentally only represents extreme opinions. This is why the analysis has to measure the token-frequency in a *relative* way. Potts and Schwarz execute this by shifting from frequency to logistic odds. This

[25] Potts and Schwarz, "Exclamatives and heightened emotion," 17.
[26] The corpus can be found on: http://semanticsarchive.net/Archive/jQ0ZGZiM/readme.html.
[27] For a detailed overview of the corpus, see Appendix A of Potts and Schwarz, "Exclamatives and heightened emotion."

happens in two steps: first, they calculate the odds of a token occurring in a certain rating category with the following formula (f1):[28]

(def. 1) $\text{count}(x_n, R) \stackrel{def}{=}$ the number of tokens of x_n in documents with rating R

(def. 2) $\text{count}_n(R) \stackrel{def}{=}$ the number of tokens of word sequences of length n in documents in rating category R (i.e., $\sum_{x_n} \text{count}(x_n, R)$).

(f1: odds) $$\text{odds}(x_n, R) \stackrel{def}{=} \frac{\text{count}(x_n, R)}{\text{count}_n(R) - \text{count}(x_n, R)}$$

We can now calculate the odds for a specific token occurring in a rating category. However, this is not yet relative enough. Potts & Schwarz shift these odds to log-odds, so it is possible to compare the *relative difference in odds* between tokens. The use of log-odds is important, because the odds could produce a distorted image. Given a token x, say the probability of occurrence in rating category 2 is 0.01 and the probability of occurrence in rating category 3 is 0.02. There is a difference in probability of 0.01. For that same token x, the probability of occurrence in rating category 4 is 0.43, and the probability of occurrence in rating category 5 is 0.44. This is also a difference in probability of 0.01. However, the odds of x occurring in rating category 3 is *twice as large* as it occurring in rating category 2, while this is far from the case for the odds in the combination category 4 and category 5. Log-odds stretch out these relative differences, and thus give a reliable comparison between a token in different rating categories. A second reason for using log-odds, is that this enables Potts and Schwarz to use logistic regression later on in their method, which is of fundamental importance for the method they eventually use to model their results.

Potts and Schwarz shift to the log-odds by using the following formula:[29]

(f2) $$\text{log-odds}(x_n, R) \stackrel{def}{=} \ln(\text{odds}(x_n, R))$$

The next step is to execute the experiment and analyse the results. Potts & Schwarz put the log-odds formula into R, which enables them to both calculate and visualize the results.[30] The R -script takes a token (including its attached values) out of the corpus, and calculate the log-odds for every rating category the token appears in. It then draws a scatterplot, each point representing the log-odds for the token appearing in that rating category, as shown in figure 1:[31]

[28] Potts and Schwarz, "Exclamatives and heightened emotion," 9-10.
[29] Ibid., 10.
[30] See appendix B for the R-script.
[31] Potts and Schwarz, "Exclamatives and heightened emotion," 9.

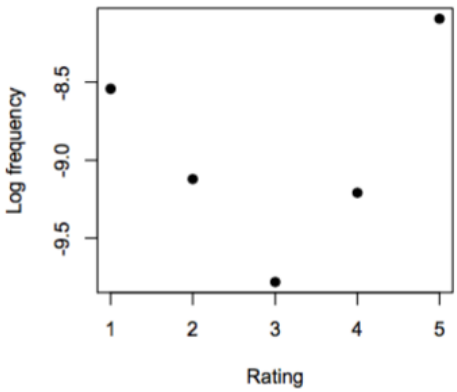

fig. 2: scatterplot of 'what a', from Potts and Schwarz (2008), p. 9.

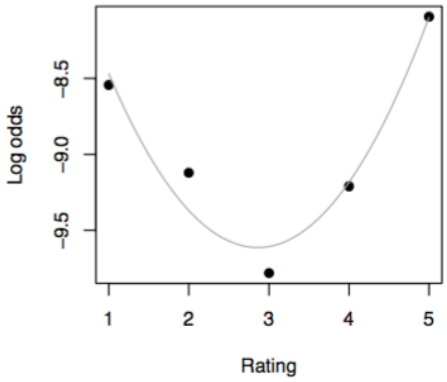

fig. 1: regression line of 'what a', from Potts and Schwarz (2008), p. 13.

To visualize the distribution of the token even more, it then draws a quadratic logistic regression line, as shown in figure 2.[32] Basically, this line represents the points that are as close to all the five data points as possible. This has to be a *quadratic* logistic regression line, because the relationship between the data points is not linear. Potts and Schwarz use *logistic* regression because this is most commonly used for data sets that have one *continuous* predictor (the rating scale, which is always the same) and one binomial *dependent* variable (the log-odds). This last variable is *binomial*, because when starting the analysis, the code interpreter has to determine for each token in the corpus if it is the specific token it is looking for, yes (1) or no (0).

The standard formula for a quadratic logistic regression line is as follows:[33]

(f3) $$y = ß_0 + ß_1x + ß_2x^2$$

In this case, the first coefficient (usually called the *intercept*) is not relevant, it only determines the value of y when x = 0, i.e. the height of the line; the experiment however, focuses only on the *shape* of the regression line.

The second coefficient, the *linear* coefficient, determines the rate at which the curve increases – it determines the slope of the line.[34,35] The third coefficient, the *quadratic* coefficient, determines the narrow- or wideness of the curve of the regression line: the bigger $ß_2$ is, the narrower the curve is – as $ß_2$ can also be negative, 'bigger' is in this sense 'further away from zero'. This is shown in figure 3, where the formula with a bigger $ß_2$ has a narrower shape than the formula with a smaller $ß_2$.

[32] Potts and Schwarz, "Exclamatives and heightened emotion," 13.
[33] Ibid., 18.
[34] Kaitlin Spooner, "Exploration of Quadratic Functions," http://jwilson.coe.uga.edu/EMAT6680Su10/Spooner/Assignment2KS/Assignment2KS.html (retrieved June 8th, 2016).
[35] Potts and Schwarz, "Exclamatives and heightened emotion," 18-19.

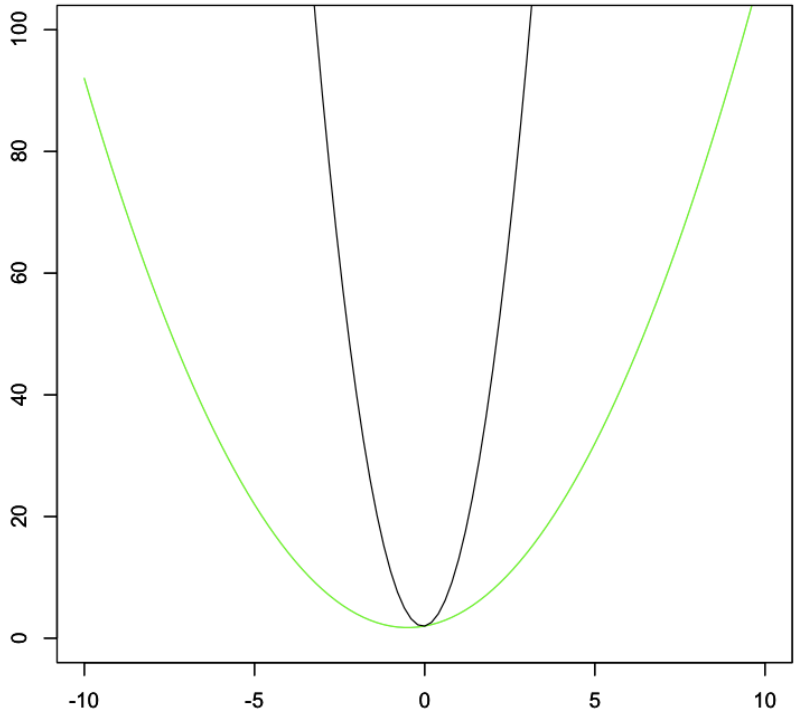

fig. 3: the green curve expresses the formula $y = 2 + x + x^2$. *The black curve expresses the formula* $y = 2 + x + 10x^2$.

$ß_2$ also determines if the curve is facing up (a U-shape), when $ß_2$ is positive, or down (a Turned-U shape), when it is negative. In conclusion, these coefficients can tell us a lot about the shape of the regression line. Figure 2 shows an example of a logistic regression line, in this case for the token 'what a'.[36]

Each phrase that exists in the corpus thus has a corresponding regression line and two attached coefficients.

For the whole analysis, Potts and Schwarz set the significance level at $p < 0.001$. This is fairly low, as the standard significance level is usually set at $p < 0.01$. An explanation for this could be that this is the kind of data that is either extremely significant or not significant at all. Some test-runs confirm this: both the linear and quadratic significance levels are often either smaller than 0.001 or bigger than 0.05.

Realization of experiment

Potts and Schwarz then run the mentioned logistic regression on their corpus, and analyse the data-output of the R-script (i.e., the regression line and the two coefficients $ß_1$ and $ß_2$). As a result, Potts and Schwarz propose three statistical profiles for exclamative phrases (uni- and bigrams) of written language, with a fourth profile for a phrase that is explicitly neutral in emotion:

[36] $ß_1$ (linear coefficient) = -1.891, $ß_2$ (quadratic coefficient) = 0.331.

Identifying distribution-types with a logistic regression model

Shape	Quadratic coef	Quadratic p	Linear coef	Linear p
U	positive	significant	–	nonsignificant
Turned-U	negative	significant	–	nonsignificant
J	positive	significant	positive	significant
Reverse-J	positive	significant	negative	significant

Table 1: statistical profiles according to Potts & Schwarz (table from Potts and Schwarz (2008), p. 21). The significance rate is set at $p < 0.001$ for all statistical profiles.[37]

Each row stands for a specific statistical profile. A visual representation of these profiles can be found on the next page (figure 4-7). The first column of table 1 sums up the possible shapes of the logistic regression line that belongs to a certain phrase (for example, the line in figure 2). A U-shape in such a plot conveys the presence of some heightened emotion with the use of that particular word or phrase, either positive or negative, as the word or phrase occurs relatively the most in extreme reviews. An example of such a phrase is 'what a' (see figure 2).

A J-shape conveys the presence of an exclamative with a positive bias, as such an exclamative occurs relatively the most in the positive end of the rating-spectrum (i.e., in four or five star ratings). A reverse-J-shape conveys the opposite, an exclamative with a negative bias, as the corresponding word or phrase relatively occurs the most in the negative end of the rating-spectrum (i.e., in one or two star reviews).

The turned-U shape is the only one in the series of statistical profiles that explicitly conveys the *absence* of an exclamative. It shows that the corresponding word or phrase occurs relatively the most in the moderate rating-spectrum (i.e., around three stars). This in turn generally means that the author does not wish to express any 'extreme' emotions.

[37] Potts and Schwarz, "Exclamatives and heightened emotion," 21.

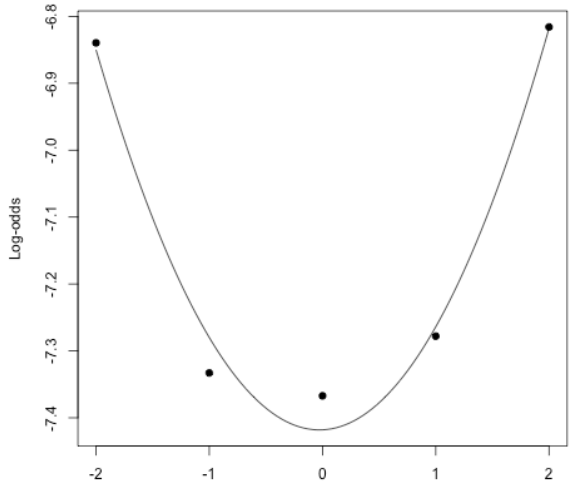

fig. 4: U-shape. Quadratic coef = 0.146, quadratic p = 0, linear p = 0.779.

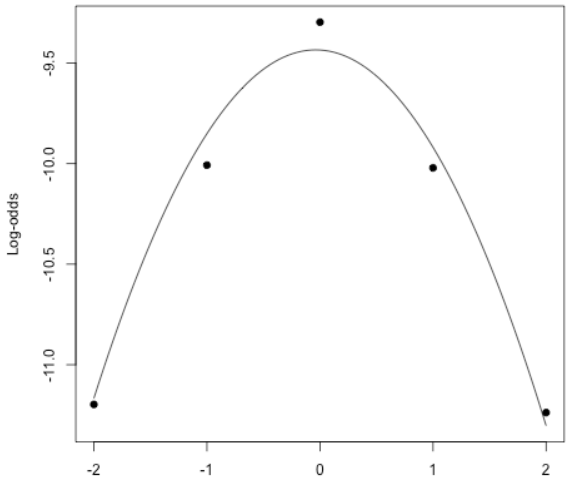

fig. 5: Turned-U shape. Quadratic coef = -0.45, quadratic p = 0, linear p = 0.850.

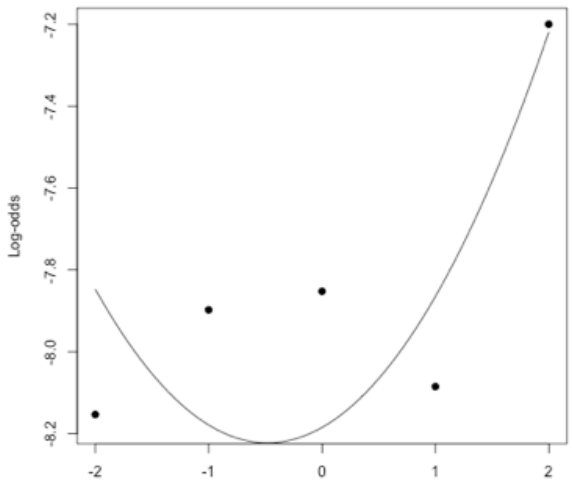

fig. 6: J-shape. Quadratic coef. = 0.163, quadratic p = 0, linear coef. = 0.157, linear p = 0.0003.

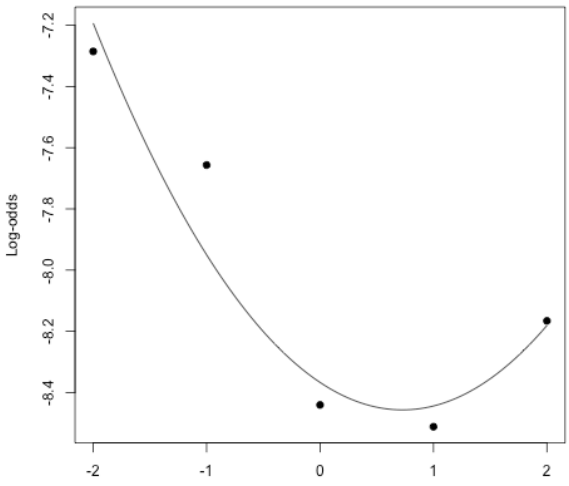

fig. 7: Reverse-J shape. Quadratic coef. = 0.17, quadratic p = 0, linear coef. = -0.247, linear p =0.

Each statistical profile then has another four aspects, aside from the shape. These aspects are represented in the next four columns of table 1. Each statistical profile includes a fixed combination of four values: the quadratic coefficient ($ß_2$, see formula 3), the quadratic p, the linear coefficient ($ß_1$, see formula 3) and the linear p. For a phrase to have the statistical profile of a J-shape, for example, both its quadratic coefficient and the linear coefficient (see formula 3) have to be positive, and have to have significant values ($p < 0.001$).

Potts and Schwarz use this set of statistical profiles to analyse their data-set by filtering out those words or phrases that distribute one of the four significant shapes, finding such words or phrases by searching for the right combination of coefficients. These statistical profiles indeed turn out to be reliable to detect exclamatives: the phrases that convey one of the significant exclamative profiles (the U-, J- or Reverse-J-shapes) are indeed likely to be exclamatives (for a list, see table 3 in chapter 3). This is a promising step towards the characterization of exclamatives, and finding unique aspects of them. Potts and Schwarz might have found a set of specific aspects belonging to exclamatives, namely the statistical profiles. This could offer more insight in how and when exclamatives are used, which in turn offers a deeper understanding of the character of exclamatives.

As the theory about statistical profiles for exclamatives seems promising, I will call this from now on the 'statistical-profiles-hypothesis': positively and negatively biased exclamatives convey respectively a J- and reverse-J-shaped statistical profile, exclamatives that can express both a positive and negative emotion convey a U-shaped statistical profile, and expressions that do not express exclamativity convey a Turned-U-shaped statistical profile. This hypothesis is compatible with hypothesis 3 (h3) in chapter 2 on page 10, as the three statistical profiles that express heightened emotion have their highest points in extreme rating categories (1 and 5 stars), while the statistical profile that does not express heightened emotion, has their lowest points in these rating categories.

Although the initial results "seem promising", as they indeed find phrases that are intuitively exclamative, there has not yet been enough testing to accept this theory of statistical profiles as fully reliable, in any case not across different languages and corpora. Potts and Schwarz themselves suggest to "gauge the success of [the statistical-profiles-hypothesis] by appeal to intuitions or by engaging in further experiments to see whether speakers genuinely regard these phrases as signals that the speaker is in a heightened emotional state."[38] I hope to contribute to the raising of the reliability of the hypothesis by testing it on a Dutch corpus in chapter 3 and 4.

Obstacles

The statistic profiles theory also brings along some problems. Potts & Schwarz mention two main obstacles:

[38] Potts and Schwarz, "Exclamatives and heightened emotion," 21.

(1) "One clearly problematic class of items consists of function words, such as my, i and perhaps this."[39] The problem is that, apart from words and phrases with an exclamative function, a lot of function words also distribute significant shapes. A proposed explanation for this is that function words occur a lot in language, and will thus almost always convey *some* significant profile. In logistic regression, even a very shallow U-shape can be significant if the phrase is frequent enough. One way to eradicate this difficulty is proposed, namely by involving the 'pureness' of the shapes. The larger the quadratic coefficient, the narrower and thus the purer the U-shape becomes. An example is shown in figure 8: the shape of the phrase 'my' is almost a straight line, while the shape of the phrase '!!' has got a narrower and more round shape. A solution might thus be to set a certain quadratic coefficient threshold for a shape to come up as a real exclamative, so it is not only dependent on the frequency and the positive or negative value of it.

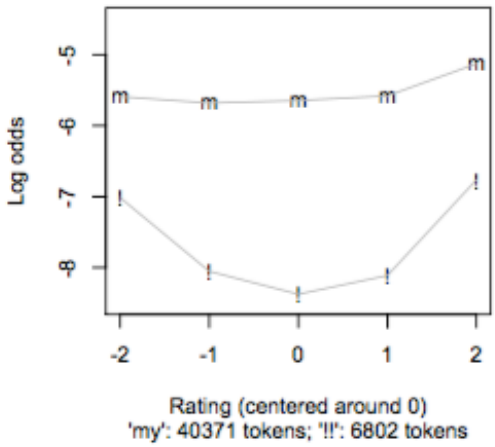

fig. 8: shallow versus narrow shapes (Potts and Schwarz (2008), p. 23).

Another proposed solution is to "simply exclude function words"[40] as a whole category, which is more radical than the first solution. Before taking such a step, I do think the nature of exclamatives has to be thoroughly investigated before excluding a whole word class out of the scope of exclamativity.

(2) The second problem is that the resulting set of exclamatives "also seem[s] to contain a subclass of elements that harbor exclamativity [...] in the context of the domain of the reviews, but not more generally."[41] This is a more or less obvious consequence of analysing data from one specific domain (e.g. book reviews), and would quite easily be solved by incorporating multiple corpora from different domains (as proposed by Potts and Schwarz). However, as I will discuss in chapter 4, this problem could also have an advantage for some specific application of sentiment analysis.

Apart from these two problem Potts and Schwarz address, there is another problem that announces itself when performing experiments on large corpora: the problem of ambiguity and synonymy.

One of the problems that occur in linguistic research, especially in these kind of corpus studies, is the fact that some words are ambiguous. This is not limited to the semantic level: some words can occur in two or more word classes, for example: 'mind' can be a verb as in "I would mind", or a noun as in "in my mind". This is a problem that might at least for

[39] Potts and Schwarz, "Exclamatives and heightened emotion," 21.
[40] Ibid.
[41] Ibid., 23.

some words influence the extracted data, as the verb ‘mind’ might have a different distribution pattern that the noun ‘mind’.

One solution would be to tag the words, but there is not yet a proper and elegant tagger on the market that will tag all of our words perfectly. This will also add some noise, and it is questionable whether the tagger-noise will be favourable to the ambiguity-noise. A solution would be to do this by hand, but as most linguistics work with large corpora this would be an exhausting and highly time-consuming task.

Another problem is that some words have synonyms, which might influence the extremeness of the distributional pattern: people who use the world ‘horrible’ usually wish to convey the same (negative) emotion as the people who use the word ‘awful’. However, this same emotion results into two different distributional patterns. The word ‘awful’ might be less used over all, so that is has a shallower regression line, while it actually conveys the same level of emotion as ‘horrible’, which might in turn have a less shallow regression line. This could slightly influence the results. However, my expectation is that words that convey the same meaning will also convey the same emotional load, will be used relatively the same (e.g. five times as much in 1-star reviews as in 5-star reviews), and thus have a similar or even the same statistical profile.

Potts and Schwarz do mention a positive result, namely the category of “items that assist in conveying exclamativity without fully determining it,”[42] such as the word ‘ever’, and the notable role that scalar-endpoint items and superlatives play in expressing exclamativity. This might also help in the polarity determining when studying exclamatives, a subject that could be the next fundamental step in the characterizing of exclamativity. This result is something that will also be shown in the next chapter, when discussing the results of the bol.com experiment.

Apart from the two main difficulties, Potts and Schwarz do propose a rather promising hypothesis that also yields results that “illuminate the nature of exclamativity and also phenomena that are conceptually and linguistically related to it.”[43] Considering Potts and Schwarz got to these results using only two corpora in two domains, only involving uni- and bi-grams, it is in conclusion a promising step in the field of pragmatic exclamativity analysis.

[42] Potts and Schwarz, “Exclamatives and heightened emotion,” 23.
[43] Ibid.

Chapter 3: The Bol.com Experiment Part 1: Comparison

In the previous chapter, I have thoroughly discussed the method of Potts and Schwarz regarding characterizing exclamatives. I have addressed the goals, the relevance of their method in the context of sentiment analysis, and the experiment they performed. The latter seems promising, but has only been tested across a specific English corpus. However, regarding the nature of the method, this experimental method should yield the same results across different corpora and different languages. This is a hypothesis I am going to investigate, by reproducing their experiment on a Dutch corpus. If the results are indeed similar to Potts and Schwarz's results, the statistical-profile-hypothesis will be somewhat more reliable, which is again a promising step in the study of exclamatives.

3.1 The Bol.com-corpus

For testing the statistical-profiles-hypothesis, I gathered a corpus that closely resembles the one that was used conducting their experiments.[44] My corpus, the *bol.com-corpus*, exists of product-reviews from the site *bol.com*, which is the Dutch equivalent to Amazon - this results in almost exclusively Dutch reviews. The product range varies from, but is not limited to, books, movies, kitchen equipment, electronic devices and furniture. The information I used for the experiment consists of 27.583 different reviews for 26.906 different products, containing 1.726.250 words in total. The connected rating scale goes from 1 to 5 stars. Table 2 shows the amount of reviews per rating category in both the Potts and Schwarz-corpus and the bol.com-corpus:

Rating	Potts and Schwarz-corpus	Bol.com-corpus
1	6219	1141
2	6817	1145
3	8942	2437
4	24402	6834
5	57402	16026

Table 2: amount of reviews per rating category.

The corpus is free of duplicate reviews or reviews without rating and vice versa, and has been stripped from capital letters, accent- and punctuation marks. The resulting set of tokens contains unigrams, bigrams and trigrams. The bol.com-corpus includes the exact same four variables as the Potts and Schwarz corpus, as is shown in the following example of a line out of the bol.com-corpus:

Word	*Rating*	*Tokencount*	*Widecount*
'aandacht'	1	27	72991

[44] See appendix C for a detailed description of the construction of the corpus.

There are some small differences between the Potts and Schwarz corpus and the bol.com-corpus.

(1) Whereas the former includes punctuation marks and accent marks, the latter does not. As the only significant punctuation mark that has been found in the Potts and Schwarz study is the exclamation mark, I do not consider the lack of punctuation marks in the bol.com-corpus a real shortcoming.
(2) The Potts and Schwarz corpus consists of metadata from two websites (Amazon.com and Tripadvisor.com), which both consist out of two separate components: a review and a summary of the review. The bol.com-corpus consists of metadata from only one website (bol.com), and only contains reviews. I have chosen not to use the summaries, as I found that the summaries from bol.com were relatively short (an average of 3 words opposed to 6 words in the Potts and Schwarz corpus) and were mostly the title of the product or simply empty: the summaries thus seemed not to be relevant. Furthermore, I do not consider the use of only one website a shortcoming, as the reviews cover a large number of products and thus also a large audience.
(3) The Potts and Schwarz corpus has considerably more reviews per item: on average 153 per book on Amazon.com and 110 per hotel on Tripadvisor.com. In the bol.com-corpus, there is an average a little over 1 review per item. I do not consider this a shortcoming, as the particular *subject* of the exclamativity should not matter, as long as the exclamativity is expressed in the same form (in this case in the form of a written review).
(4) The Potts and Schwarz corpus contains reviews of only two types of items: books and hotels. The bol.com-corpus contains probably significantly more types of products, as I have not picked the reviews on the basis of the product types they were about. Although as said in (5), I do think this should not matter because of the same form of exclamativity, I do think there are some differences to be noted in the results. Whereas in the Potts and Schwarz results there will be significant words as 'bed', 'reception', 'book' and the like, related to either hotels or books, in the bol.com results will be seemingly more arbitrary words. This is not a shortcoming of the experiment, but something to take into account when discussing the results.
(5) The Potts and Schwarz corpus is approximately four times as large as the bol.com-corpus (103,782 against 27,583 reviews). This could translate into less significant results or even no significant results at all when it comes to weaker, which is definitely something to take into account. However, in other similar studies, analyses run on smaller corpora like the Chinese MyPrice corpus with 17,513 reviews.[45] In conclusion, the bol.com-corpus size might influence some of the results (especially when comparing it to the four times as large Potts and Schwarz results), but does not lack significance at all.

3.2 Method

The method I will be using to recreate Potts and Schwarz's experiment, is roughly the same as explained in chapter 2. I will put phrases from the bol.com corpus into the in chapter 2

[45] Noah Constant et al., *UMass Amherst Linguistics Sentiment Corpora,* last updated in January 2009, http://semanticsarchive.net/Archive/jQ0ZGZiM/readme.html (retrieved May 20th, 2016).

discussed R-script, and will get an output of a scatter- and regression line-plot and the relating values of the quadratic and linear coefficients and significance scores.

For the goal of recreating the experiment as accurately as possible, I used the R-script Potts and Schwarz developed to test their corpus as a basis. For further analysis, I added functions to extract tokens with significant statistical profiles.[46]

One thing that is different from the examples I used in chapter 2, is the rating scale I will be using. Just as Potts and Schwarz propose later in their experiment.[47] For pure convenient reasons, I shift the rating scale from 1 to 5, to -2 to 2. This is because the pure U-shapes will then be easier to recognize, as the linear coefficient now has to be close to zero and thus insignificant: when $ß_1$ is zero, the counterpoint for the quadratic line is when x = 0. When using a rating scale from -2 to 2, this counterpoint will be in the exact middle of the chart and thus will the U-shape be perfectly symmetric and pure. Figure 9 and 10 show this shift of the rating scale:

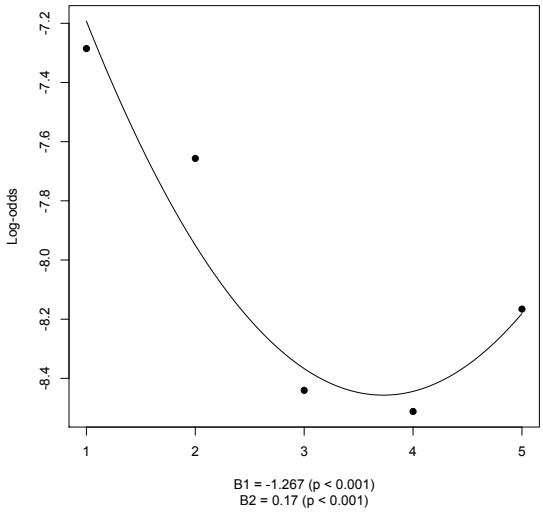

fig. 9: 'absoluut' on a 1 to 5 rating scale

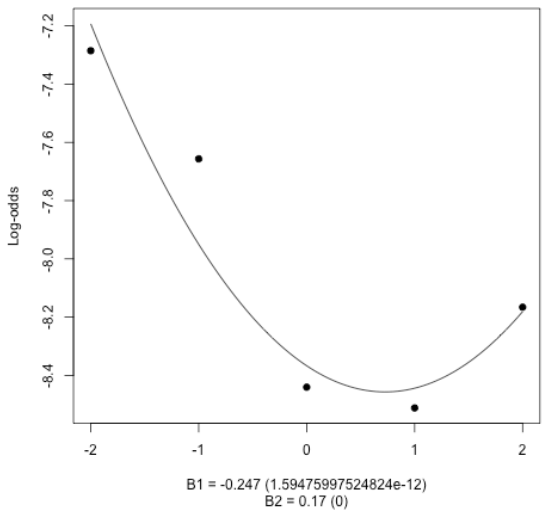

fig. 10: absoluut on a -2 to 2 rating scale.

The significance level of the conducted analysis is set at the same value as in the Potts and Schwarz experiment: $p < 0.001$ for both the quadratic and the linear coefficient. Therefore, in the following results and discussion, 'significant' will mean that the corresponding $p < 0.001$ and non-significant will mean $p >= 0.001$.

My experiment to test the statistical-profiles-hypothesis of Potts and Schwarz consists out of two parts.

For the first part, I will use an important part of the results published in the Potts and Schwarz paper and compare them with my own results. These specific Potts and Schwarz results consist out of a list of lemmas that produces one of the three statistical profiles for exclamatives. I will translate these lemmas to Dutch, and run the logistic regression analysis

[46] See appendix B for R-script.
[47] Potts and Schwarz, "Exclamatives and heightened emotion," 19.

on them.[48] I will then check for each English lemma if the Dutch counterpart(s) has or have a significant statistical profile as well. If the statistical-profiles-hypothesis is correct, and thus consistent across multiple languages, the Dutch counterparts should have significant statistical profiles as well. In this part, I will also address the obstacles that Potts and Schwarz pointed out, and discuss if and how this affects my experiment as well.

For the second part, I will run the logistic regression on the bol.com-corpus and filter out all of the significant statistical shapes, after which I will discuss the results with respect to the meaning of the statistical profiles. This final part, including the hypothesis for it, can be found in chapter 3, in which I will elaborate on this specific part of the experiment.

I will assume that the statistical-profiles-hypothesis is correct, and therefore expect the results of the bol.com corpus to over all be the same as the results of the Potts and Schwarz corpus, thereby underlining Potts and Schwarz's prediction that the statistical profiles-approach works across different corpora and languages.

For part one, I expect the majority of the English lemmas to have a Dutch significant counterpart. There could be exceptions, because there are always differences in nuance and meaning between two different languages, but that should not affect the experiment too much, as English and Dutch are from the same linguistic family, the (West-)Germanic family.[49]

3.3 Results and Discussion

In table 3, the lemmas published in the Potts and Schwarz experiment are contrasted with the corresponding lemmas in the bol.com-corpus experiment. The first column consists of the lemmas from the Potts and Schwarz experiment that have a significant U-, J- or Reverse-J-shaped statistical profile. **Bold** tokens are lemmas "whose shapes are limited to U, J, and Reverse-J for all [...] corpora"; *italic* lemmas are lemmas that cannot translate to a lemma in the bol.com corpus (caused by the lack of punctuation marks in the latter); regular formatted lemmas are lemmas "whose shapes are limited to U, J, and Reverse-J in [75% of the] corpora." Potts and Schwarz unfortunately did not publish the shape and coefficient values for each token. This may be because each lemma appears in three or four corpora, which yields at least three sets of shape and coefficients. Taking the mean of them might influence the accuracy of the study; taking all of them only causes a cluttered overview which is hard to compare with the results of the bol.com corpus.

The second column exists out of the Dutch counterpart lemmas, the next three columns are the corresponding coefficient values and shape for that Dutch lemma. All of the Dutch lemmas have a significant statistical profile. The second column sometimes reads "no significant results": this means that for the thereafter mentioned Dutch lemma's, there was no result with quadratic $p < 0.001$. This could either be because there simply weren't enough tokens of that lemma, or because the distribution of this lemma across the rating categories

[48] For the translation of the significant tokens presented in the first column of table 3, I used the Van Dale English to Dutch dictionary and maintained the first translation. If there was more than one translation, e.g. 'I could' could mean 'ik zou' or 'ik kon', I used the first (up to) three options or chose one that was clearly relevant in the context of reviews.

[49] Anne E. Baker et al., *Taal en Taalwetenschap* (West-Sussex: Wiley-Blackwell Publishing, 2013), 254-258.

was too scattered to convey a significant shape. Unfortunately, it is not possible to compare the shapes of each pair of lemmas, for the mentioned reason that Potts and Schwarz did not publish the shape and coefficient values for each token.

English lemmas	Dutch lemmas	Linear coefficient	Quadratic coefficient	Shape
!	-	-	-	-
!!	-	-	-	-
absolute	absoluut, absolute	-0.247 0.293	0.17 0.246	Reverse-J U
absolutely	absoluut	-0.247	0.17	Reverse-J
again !	-	-	-	-
all	alle	0.051	0.056	U
am	ben	0.051	0.074	U
any	(no significant results for ‘enige’, ‘enkele’, ’wat’)	-	-	-
anyone	(no significant results for ‘iemand’, ‘wie dan ook’)	-	-	-
best	beste	0.157	0.163	J
book	boek	-0.013	0.025	U
couldn’t	(no significant results for ‘kon niet’, ‘konden niet’)	-	-	-
even	zelfs	-0.085	0.108	U
ever	ooit	0.163	0.248	U
ever !	-	-	-	-
ever had	ooit heb	0.115	0.567	U
every	alle	0.051	0.056	U
have ever	ooit heb	0.115	0.567	U
i	ik	-0.075	0.051	Reverse-J
i am	ik ben	0.063	0.099	U
i could	(no significant results for ‘ik kon’, ‘ik zou’)	-	-	-
i have	ik heb	-0.029	0.125	U
i’ve	ik heb	-0.029	0.125	U
i’ve ever	ik ooit heb	0.089	0.633	U

is the	is het	-0.007	-0.063	Turned-U
it !	-	-	-	-
life	(no significant results for 'leven')	-	-	-
must	moeten	-0.13	0.09	Reverse-J
my	mijn	-0.004	0.099	U
never	nooit	-0.028	0.157	U
new	(no significant results for 'nieuw', 'nieuwe')	-	-	-
one of	(no significant results for 'een van')	-	-	-
simply	gewoon	-0.024	0.116	U
the best	de beste	0.239	0.196	J
this	deze dit	0.004 -0.048	0.054 0.071	U Reverse-J
this is	dit is	0.008	0.146	U
time	(no significant results for 'tijd')	-	-	-
what	(no significant results for 'wat')	-	-	-
what a	wat een	-0.071	0.352	U
will	(no significant results for 'zal', 'zult', 'zullen')	-	-	-
will never	(no significant results for 'zal nooit', 'zult nooit', 'zullen nooit')	-	-	-
wow	(no significant results for 'wow', 'wauw')	-	-	-
wow !	-	-	-	-

Table 3: listing of corresponding English and Dutch lemmas.

For *all* bold lemmas, i.e. the lemmas that have a significant shape in each of the Potts and Schwarz corpora, there are significant corresponding Dutch lemmas to be found (Dutch counterpart lemmas that have a significant profile as well). This is a first promising result for

the affirmation of the hypothesis. As for the 25 'regular' lemmas, i.e. significant in 75% of the Potts and Schwarz corpora, there are 13 corresponding Dutch significant lemmas. This is less promising: an accuracy rate of little over 50 percent will not affirm the hypothesis.

There could be two reasons for this result. The English lemmas that do not have a Dutch counterpart are fairly grammatically scattered – the set involves nouns, verbs, determiners, et cetera. However, the English lemmas that dó have a Dutch counterpart, mostly seem to convey a more extreme meaning, for example: 'absolutely', 'all', 'best', 'ever', 'the best' and 'what a' intuitively convey more exclamativity than e.g. 'i could', 'life', 'new', 'one of', 'time' or 'what' do. Another reason for the low accuracy could be that the bol.com corpus is relatively small compared to the Potts and Schwarz corpus. Considering the fact that the 'regular' lemmas did not show significant shapes in *all* four of their corpora, it could just be that these lemmas are not used that often, and quickly fall out of scope in a relative small corpus like the bol.com-corpus.

Conclusion

After the comparison of the results for the Dutch equivalent of the significant English tokens, I want to conclude that the statistical-profiles-hypothesis yields the same results for Dutch as for English. This implies that the hypothesis indeed yields across different languages, at least for Germanic languages. There are some tokens that are not significant in Dutch corpus whereas they are in the English corpus, but I want to argue that this is probably because of the relative small size of the Dutch corpus, and the difference in nuance when expressing exclamatives in English and Dutch. It is however clear that the most important tokens do have a significant Dutch counterpart, from which I conclude that Potts and Schwarz were right in assuming that there hypothesis holds for languages different than English as well.

Chapter 4: Bol.com Experiment Part 2: More Results

In the second part of my experiment, I will test the statistical-profiles-hypothesis somewhat more thoroughly. Side-stepping from specific results out of the Potts and Schwarz paper, I will filter out all the phrases with significant statistical profiles and judge them on exclamativity, positive and negative bias. This will give an insight in the accuracy of the statistical-profiles-hypothesis, not only in general but also specifically for Dutch. As a second aspect of this chapter, I will address the by Potts and Schwarz raised difficulties, and explore if these are obstacles for the bol.com experiment as well.

4.1 Hypothesis

I want to argue that there are intuitions regarding words and what kind of emotion (or no emotion at all) these words express (e.g., the word "amazing" intuitively expresses a positive emotion). My hypothesis is that for the Dutch language, at least part of these intuitions can be confirmed using the statistical-profiles-hypothesis. I expect different results for each of the statistical profiles. Regarding the U-shaped profile, presumably conveying exclamatives without a polarity bias, I expect the results to contain maximizing intensifiers like 'zeer', 'heel erg' and 'enorm', as well as specific words with a fixed exclamative value like 'wow' and 'jeetje'. These are all words that convey an extreme meaning, while not having a fixed polarity – these are therefore likely to be used in extreme rating categories (1- and 5-star reviews). For example, the word 'zeer' could be used in a sentence that expresses negativity ("Dit product is zeer slecht!") as well as in a sentence that expresses positivity ("Het product beviel me zeer goed!").

As for the J-shaped profile, which denotes exclamatives that are positively biased, I expect positively biased exclamative tokens such as "geweldig", "top" and "zeer goed". Of course the opposite is expected for the Reverse-J-shaped profile, denoting negatively biased tokens such as "slecht", "verschrikkelijk" and "niet goed".

There will also be some problems when performing the experiment. Acknowledging the first problem Potts and Schwarz raise, the high presence of function words, I expect there to be a number of function words with significant statistical profiles. This problem might not limit itself to function words – there might be a number of seemingly arbitrary tokens that convey a significant shape. According to Potts and Schwarz, these tokens can presumably be found by checking the mentioned 'pureness' of the particular shapes, i.e. eliminating the shallower shapes.

As for the second problem, the high presence of domain-specific words, I do not expect the context of the domain to result into significant exclamative tokens that are only exclamative in that context. This is because the bol.com-corpus contains reviews about a large set of different products.

4.2 Discussion

The complete list of results of the logistic analysis can be found in appendix A.

The hypothesis expressed the prediction that the phrases with a U-shaped profile would contain both intensifiers and tokens with a set exclamative value. There are indeed some intensifiers to be found in the U-shaped results: 'enorm', 'heel erg', 'hele', 'helemaal', 'mooiste', 'ontzettend' and 'vet'. This is a rather small set of intensifiers, missing some obvious intensifiers like 'zeer' and 'heel', however, that could be the consequence of the rather small data set: the data could still be too randomly distributed. There are also some other tokens in the list that could adopt the role of intensifier, like 'echt' or 'zoveel'.

As for the tokens with a fixed exclamative value, there are less to be found: candidates are 'ik ooit heb', 'wat een' en 'zelfs', but exclamatives like 'wow' and 'mijn god' fail to occur. Again, this might be a consequence of the relatively small data set, but also of the domain of written language: people seem to use exclamatives like 'wow' and 'jeetje' more in spoken language than in written language. However, this is still mere speculation, and cannot fully account for the lack of set exclamatives found in the bol.com-corpus.
An interesting find in the U-shape-profile is the appearance of the words "groeten" and "groetjes", roughly translated to "greetings". It seems that author who have uttered a strong opinion, feel the need to explicitly state an ending of their review and (presumably) sign it with their name. This may give us insight to an aspect of exclamativity, namely the need to express an opinion in a rather completed form, while at the same time taking full responsibility for it.

Another trend in the U-shaped profiles is the explicit expression of emotions. The tokens 'hoop', 'huilen', 'mooiste', 'spijt', voel' and 'zielig' convey an explicit emotion or emotional judgement. In retrospect, it is not odd that these kind of words are significant exclamatives: the mentioned definition of an exclamative *is* the conveying of an emotion. It is valuable to see that this aspect of exclamativity is also expressed by the corpus analysis.

In conclusion, for most of the phrases with U-shaped profiles, there is a plausible explanation that can be found as to why they are apparently exclamatives in this context. Thus far it seems that the statistical-profiles-hypothesis still works for the Dutch corpus.

The phrases with J-shaped profiles seem to be even more promising. Although it is a rather small set, consisting out of only twelve tokens, the majority obviously conveys an exclamative meaning with a positive bias. The tokens like 'je' and 'we' that are not clearly positively biased, are mostly function words and indeed show a merely shallow J-shape: whereas 'fantastisch' and 'geweldig' have quadratic coefficients of 0,225 and 0,146, 'je' and 'we' have quadratic coefficients of 0,024 and 0,064.

The phrases with Reverse-J-shaped profiles are less clear. Even though there are some obvious negatively biased exclamatives like 'kinderachtig' and 'slechte', the majority of the words seems rather arbitrary. However, there is a significant part of the set that involves content words, something that could be explained by the tendency of people to explain what exact part of the product they did not like.

4.2.1 Obstacles

My hypothesis stated that the bol.com-corpus would also have to deal with the problem of the significance function words, as these are extremely frequently used. This is indeed the case: a small grasp out of the results shows us a set of function words such as ‘deze’, ‘dit’, ‘nou’, ‘zo’. Some of these function words indeed show a shallow shape when contrasted with clear exclamatives, but a striking number of them actually don’t - see figure 11 and 12. However, as said, before drawing any conclusions, there has to be a more detailed investigation on the usage of function words in the context of exclamatives, something that can unfortunately not be covered in the scope of this thesis.

Finally, the hypothesis did not foresee any problems regarding context-significance, that is, the results containing significant tokens that are only significant in the domain of the dataset. This was in fact a mistaken assumption: there is definitely a set of tokens to be found that are closely related to the domain of the corpus, such as ‘cd’, ‘boek’, ‘dvd’, ‘juf’, ‘serie’, et cetera. This can either be explained by (again) the relative small size of the corpus, or the fact that the variety of products in the dataset is still too small. However, I think this should not always have to be a problem. It could be useful when someone wants to know what kind of products or subjects are being reacted to in terms of more extreme emotions (U-shaped-profile), negative or positive responses (J- and Reverse-J-shaped profiles) or mediocre responses (Turned-U-shaped profiles). This aspect of domain-specific significant tokens could thus both be problematic (in the context of the general study on exclamatives) and helpful (in a more commercial, specific context).

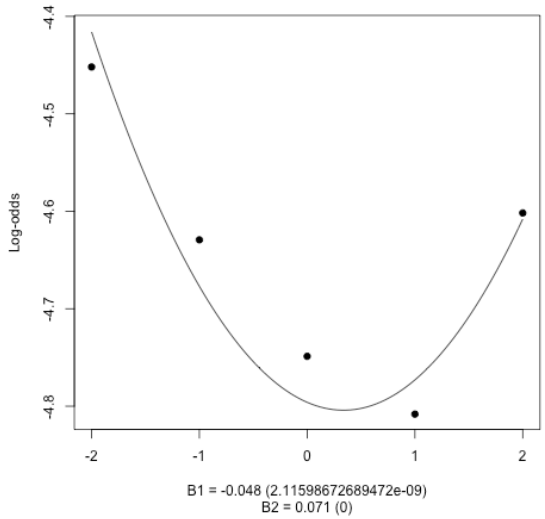

fig. 11: logistic regression line of the token ‘dit’. B1 is the linear coefficient, B2 the quadratic coefficient.

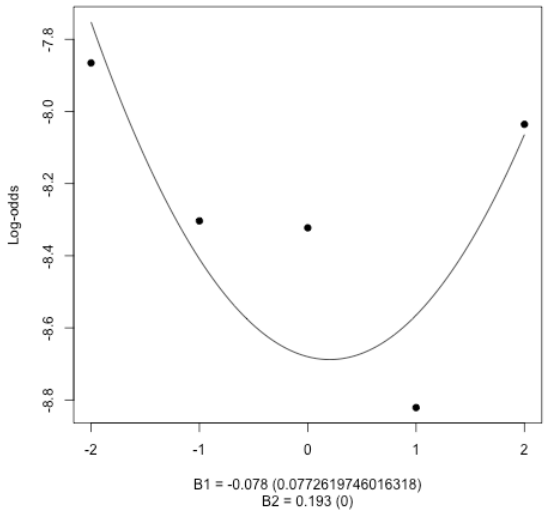

fig. 12: logistic regression line of the token enorm’.

4.2.2 The Turned-U Statistical Profile

I now want to turn to a short analysis of the Turned-U-shaped profile, a profile that Potts and Schwarz have explicitly chosen not to discuss in their paper, but do seem to think of as a valuable part of sentiment analysis.[50]

Firstly, there are a number of concessive conjunctions that convey a significant Turned-U shape. This is to be expected, as a Turned-U shape conveys moderate reviews (2, 3 or 4 stars) and comparing conjunctions are used to make some consideration on the pros and the cons of a product. A moderate review obviously has to maintain both pros and cons, conveying that the author does not have any heightened emotional feelings about the product, but is more or less satisfied with the product. The tokens found in this context are the following: 'alhoewel', 'desalniettemin', 'desondanks', 'hoewel', 'maar toch', 'ondanks', 'toch' (roughly translated to 'although', 'nevertheless', 'in spite of', 'even though', 'but still', 'yet'). This first find is thus in line with intuitive aspects of the Turned-U shape.

A token-class that might not be intuitively present, is the class of names. Interestingly enough, close to seven percent of the results were in fact names (either names of persons or titles of books or movies). This leads to another remarkable find, namely that a lot of tokens in the Turned-U results seem to indicate that a lot of authors in the moderate rating-spectrum tried to explain their opinions in quite some details, especially when stories (in either books or movies) were involved. This is derived from the fact that a lot of results were words about syntactic parts of stories ('afwisseling', 'gebeurtenis', 'hoofdpersonage', 'hoofdstukken', 'personages', 'perspectief', 'verhaallijnen'), as well as a set of functional conjunctions that seem to indicate that the author tries to express a well set-up story (e.g. 'daardoor', 'hierdoor', 'immers', 'tussendoor', 'uiteindelijk', 'einde'). These observations lead to the following hypothesis I want to propose:

> (4) Authors of a fairly moderate opinion that is by no means extreme, and therefore does not have notable (extra) exclamative or emotional content, are likely to produce a well-structured and detailed argument.

Of course this is a hypothesis based on just one language and one corpus, so there is a fundamental need for investigation before there is anything definite to be said on this. However, I do think the hypothesis is in line with the results of the above experiment, and is something worth investigating as it could give an important insight in the syntactic ways of expressing sentiment, and in particular exclamativity.

4.3 A Corpus-related Problem: Noise

Before concluding this chapter, I want to address the role of unwanted noise in the analysed data, in the bol.com-corpus as well as in the Potts and Schwarz corpus.

---

[50] "For example, items with a Turned-U distribution are 'un- exclamatives' — hallmarks of balanced reasoning. This information too can be put to good use in understanding pragmatic inferences, especially those that concern the speaker's emotional state." (Potts and Schwarz, "Exclamatives and heightened emotion," 24).

The first obvious noise is the fact that most of the words and phrases in natural language are ambiguous: if not only in meaning, then also in word type (the word 'can', for example, can refer to the verb or to the noun as in 'soda can'). This distorts the results of a corpus analysis like the one in this thesis, as the algorithm takes all occurrences of a word together into one analysis. This could easily be solved by using a word-tagger before running the algorithm: the current problem however, is that there is not yet a tagger precise enough to take away the said distortion, and tagging by hand is undoable when working with such large corpora. One way to tackle this problem in the future might be to run the same algorithm on two versions of the corpora, one that is tagged and one that is not, and comparing the results. However, there is still difficulty when deciding which set of results to take, because how would one measure the amount of noise in each corpus?

A second form of noise is the fact that people make mistakes while writing. This results in grammar or spelling mistakes, which in turn results in distortion of the dataset. However, I do think this could be solved by taking a corpus large enough so that these mistakes become insignificant. Another approach would be to check each corpus on non-existent words and correcting them, but that might cause the problem of overcorrecting, for without context one can never know for sure which word the author originally meant.

A third difficulty is the fact that when people want to underline the emotional meaning of a word, they sometimes tend to stretch the word out. In the bol.com-corpus, there was not only the token 'heel', but also the tokens 'heeel', 'heeeel' and 'heeelll'. There are not enough of these to make a significant difference, either by conveying a statistical profile or diminishing the count of the 'heel' token to which they actually can be assigned, but they do get left out when searching for perhaps exclamative sentences or authors. However, I do not think this will be a real problem, because of the fundamental minority of such phrases.

The best way so far to deal with noise when analysing corpora, is to work with a corpus as large as possible. This is a relatively easy way to make the noise insignificant.

Conclusion

In conclusion, I have argued that, following the conclusion in chapter 3, the statistical-profiles-hypothesis Potts and Schwarz proposed in their paper can be confirmed as far as the Dutch language is involved. There are still some notable differences and phenomena that need to be thoroughly investigated (e.g. the occurrence of seemingly arbitrary words in Reverse-J-shapes, the problem of function words and the lack of fixed exclamatives like 'wow'), but so far, the results seem rather promising.

I also want to briefly address the relevance of the previous experiment for the research on exclamatives and sentiment analysis. I do think Potts and Schwarz have contributed to the understanding of how to denote exclamatives, as the visual aspect of their experiment is really clear and intuitive. The statistical-profiles-theory can be of value when taking the first step of sentiment analysis, namely the distinguishing between neutral linguistic items and those with sentiment. This could contribute towards the automation of sentiment denoting, especially when they run the experiment on different types of corpora: this will eliminate the

problem of the different ways of expressing sentiment across different mediums, as mentioned in chapter 1.

Apart from pure exclamativity, the J-shaped and reverse-J-shaped profiles can be of use as well: when gathering enough data to make a solid prediction of which types of phrases have either a positive or a negative bias, it can partly solve the problem of words with fixed sentimental values, as mentioned in chapter 1 ("a sentence containing sentiment words may not express any sentiment", and the other way around).[51]

Over all, I want to conclude that the statistical-profiles-hypothesis can be of great value in the process of automatizing sentiment analysis, when tested across enough different corpora.

[51] Liu, *Sentiment Analysis and Opinion Mining,* 12.

Conclusion

The question introduced in the beginning of this thesis was: what role do exclamatives play in the context of sentiment analysis, and is there an algorithmic way to detect exclamatives and categorize them?

The thesis provided an answer to the first part of the question in chapters 1 and 2. An exclamative seems to be the first step towards the process of sentiment analysis: distinguishing between neutral and exclamative items is of crucial importance. To have an algorithm that accurately distinguishes between the two items, is a big step towards automatizing the process of sentiment analysis.

As for the second part of the question, I do think Potts and Schwarz have found a promising way into the algorithmic analysis of detecting exclamatives. When analysing patterns of linguistic parts that convey the same sentiment, as the statistical-profiles-hypothesis means to do, there is a great chance of finding valuable information about the nature of exclamatives. This has been shown by the promising results both Potts and Schwarz and chapters 3 and 4 of this thesis have delivered.

Although this only involves non-polar exclamatives, the statistical profiles that are typical for biased exclamatives seem to be rather accurate as well. I would therefore encourage future researchers to test the statistical-profiles-hypothesis on more and different corpora. This should make sure that the general linguistic profiles of exclamativity are not domain-specific, as they presumably still are in the Potts and Schwarz experiment. Comparing the results from different languages can give insight to cross-lingual differences in the expression of sentiment, which is very useful information for sentiment analysis and companies that have markets across the world and strongly rely on big-data.

I also think it would be interesting to compare the results of a statistical approach with that of machine-learning approaches: it might just be the case that they complement each other on some levels, as they both rely on mere prediction when applied to sentiment analysis automation.

In conclusion, the study of the nature of exclamativity is valuable in the field of sentiment analysis. In the process of automatizing the latter, the first step still remains the detection of language expressing sentiment verses language that does not. Constructing an algorithm that fulfils the role of this detector is therefore an important first step towards the automation process of natural language processing.

Bibliography

- Baccianella, Stefano, Esuli, Andrea, and Sebastiani, Fabrizio. "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining." *Proceedings of the Seventh Conference on International Language Resources and Evaluation* (LREC'10) (2010), 2200–2204. http://zeynepaltan.info/4-SentiwordNet.pdf (retrieved June 10th, 2016).
- Baker, Anne E., Don, Jan, and Hengeveld, Kees. *Taal en Taalwetenschap*, 2nd edition. West-Sussex: Wiley-Blackwell Publishing, 2013. 1st edition 2002.
- Boon, C.A. den, and Hendrickx, R. *Dikke Van Dale Online 2015*. http://vandale.nl (retrieved May 25th, 2016).
- Chernilovskaya, Anna. *Exclamativity in discourse: Exploring the exclamative speech act from a discourse perpective*. Dissertation Utrecht University (2014).
- Constant, Noah, Davis, Christopher, Potts, Christopher, and Schwarz, Florian. *UMass Amherst Linguistics Sentiment Corpora*. Last updated in January 2009. http://semanticsarchive.net/Archive/jQ0ZGZiM/readme.html (retrieved May 20th, 2016).
- Liu, Bing. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers (2012), e-book only. http://www.morganclaypool.com/doi/abs/10.2200/S00416ED1V01Y201204HLT016.
- Pang, Bo and Lee, Lillian. "Opinion Mining and Sentiment Analysis." *Foundations and Trends in Information Retrieval* 2/1-2 (2008). 1-135. https://www.cse.iitb.ac.in/~pb/cs626-449-2009/prev-years-other-things-nlp/sentiment-analysis-opinion-mining-pang-lee-omsa-published.pdf (retrieved June 10th, 2016).
- Potts, Christopher and Schwarz, Florian. "Exclamatives and heightened emotion: Extracting pragmatic generalizations from large corpora." Ms., UMass Amherst, 2008.
- Spooner, Kaitlin. "Exploration of Quadratic Functions." http://jwilson.coe.uga.edu/EMAT6680Su10/Spooner/Assignment2KS/Assignment2KS.html (retrieved June 8th, 2016).

Appendix A: Bol.com Results

TURNED-U-SHAPE RESULTS (quadratic $p < 0.001$; quadratic coefficient < 0; linear $p > 0.001$). This is a list of uni-, bi- and/or tri-grams that all convey a significant statistical profile with the aforementioned combination of significance level and coefficients:

aangenaam
aantal
af
afwisseling
algemeen
alhoewel
allende
anton
beeld
beetje te
behoorlijk
bepaald
bepaalde
bewijzen
bond
broers
daardoor
deels
desalniettemin
desondanks
diner
doo
dorien
duurde
een aantal
einde
eliza
elkaar
enigszins
eva
gebeurtenis
gemogen
heden
hen
herman
hierdoor
hoewel
hoofdpersonage
hoofdpersonages
hoofdstukken
iets
ietwat
immers
interessant
interessante
is het
jonathan
koch
komt
krijgt
kritische
laarmans
lastig
linda
lisa
maar toch
marjolein
meestal
minder
miste
moord
moordenaar
naar
niet echt
nieuwsgierig
oeroeg
ondanks
onderhoudend
ontspannend
over
pas
personages
persoon
persoonlijk
persepctief
redelijk
roman
silvia
sommige
spectre
storend
suze
tbs
te
thema
toch
toe
toilet
tussen
tussendoor
uiteindelijk
uitgewerkt
verder
verhaallijnen
verhulst
vermakelijk
vermoord
verteld
verwerkt
vlucht
voornamelijk
vraagtekens
vragen
vrij
vrouw
wanneer
wie
wordt
zaken
ze
zich

zijn
zorgt

U-SHAPED RESULTS (quadratic $p < 0.001$; quadratic coefficient > 0; linear $p > 0.001$). This is a list of uni-, bi- and/or tri-grams that all convey a significant statistical profile with the aforementioned combination of significance level and coefficients:

absolute
alle
begonnen
ben
besteld
bestelling
boek
bol
carry
cd
com
deze
dikke
dit is
dvd
echt
enorm
eten
film
ga
ge
gekocht
gekregen
geleverd
gelezen
gewoon
gezien
gezond
gisteren
groeten
groetjes
heb
heel erg
hele
helemaal
hoop
huilen
ie
ik ben
ik heb
ik ooit heb
inmiddels
juf
jullie
kan
keer
kg
klaar
kluun
kon
kortom
laat
m
maanden
meesterwerk
meisjes
mensen
mijn
mooiste
n
nog
nooit
nu
ontzettend
ooit
ooit heb
opnieuw
raad
serie
service
slee
spijt
t
tafel
u
uitgelezen
vergeten
vet
voel
vol
volgende
vriendelijke
wachten
wat een
weken
werkelijk
wij
woord
word
x
zelfden
zelfs
zielig
zo
zoveel

J-SHAPED RESULTS (quadratic $p < 0.001$; quadratic coefficient > 0; linear $P < 0.001$; linear coefficient > 0). This is a list of uni-, bi- and/or tri-grams that all convey a significant statistical profile with the aforementioned combination of significance level and coefficients:

beste
de beste
echte
fantastisch
geweldig
geweldige
iedereen
je
super
top
we
zeer

REVERSE-J RESULTS (quadratic $p < 0.001$; quadratic coefficient > 0; linear $P < 0.001$; linear coefficient < 0). This is a list of uni-, bi- and/or tri-grams that all convey a significant statistical profile with the aforementioned combination of significance level and coefficients:

absoluut
acteurs
al
artikel
bedenken
boeken
dacht
dit
eens
gegooid
geld
ik
kinderachtig
koop
kopen
kraan
lees
moeten
niemand
niets
niks
nou
slechte
uu
vegetarisch
verkocht
verkopen
vries
zeg

Appendix B: R-script

For the most part of the experiment, I used the exact code of Potts and Schwarz (available upon request from either me or Potts and Schwarz). This code has sufficient commentary by Potts and Schwarz.

However, I implemented two rather small functions:

(1) Instead of plotting per phrase, I declared a list called 'Lijst' that holds all the phrases that need to be plotted, and made the R-script loop through this list:

```
Lijst=list("//phrase1","//phrase2")
  for(word in Lijst){
      Phrase=word
      .
      .
      .
      }
```

When, for example, "//phrase1" gets replaced by "zeer", and "//phrase2" by "erg", the script will first perform logistic regression and create a visual result (i.e. a plot with a regression line) for the phrase "zeer" – it will then do the same for the phrase "erg".

(2) When acquiring the results for chapter 4, I used follows blocks of statements to acquire precisely one of the four statistical profiles (see table 1):

```
if( (quadraticP < 0.001) & (quadraticCoef < 0) & (linearP > 0.001) )
{// plot phrase }
```

This specific block of code filters out the Turned-U shapes: the quadratic coefficient is negative and the quadratic p is significant, while the linear p is non-significant. The linear coefficient is not present in this block of code, because it is irrelevant to the Turned-U shape.

Appendix C: Bol.com Scraper

The bol.com-corpus has been built in two parts.[52] Both of these parts are executed in Python. For the complete code and files: these are available upon request and enclosed to this thesis.

The first part is to gather data from the bol.com website. This is done by (1) entering a valid product ID in the standard bol.com product link; (2) checking if there are reviews at all; (3) iterating through all the reviews, for each one downloading the product-ID, review title, review text, rating, rating range, user location, submission time and last modification time. In the experiment, I have only used the review title, review text and rating: the rest of the data has been used to check for duplicates; (4) listing the data in a .csv file.

The bol.com product-ID's seemed to be randomly generated, so I had to download these as well. This has been done by using the package BeautifulSoup to let the Python-program search on bol.com for specific, pre-determined words (book titles and commonly used Dutch words).

The second part was to process the data. After downloading the necessary data, I first checked to see if there were no duplicate reviews. Then I stripped all the review-texts from capital letters and punctuation marks. After this, I used the Python dictionary functionality to categorize all the different phrases by rating: first for unigrams, then for a list of specific bigrams (those used by Potts and Schwarz and those relevant for chapter 4), and then for a list of specific trigrams (same as for bigrams). These were put in a .csv file which I converted into a .frame file, the extension used in the R-script.

[52] I would like to express my gratitude to Jan de Mooij, who is partly responsible for the code that has been used for the bol.com scraper.