UNIVERSITEIT GENT

Faculteit Letteren en Wijsbegeerte

Tipping the scales

Exploring the added value of deep semantic processing
on readability prediction and sentiment analysis

Het effect van diepe semantische analyse
op leesbaarheidspredictie en sentimentanalyse

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in de Taalkunde aan de Universiteit Gent te verdedigen door

Orphée De Clercq

Dit onderzoek werd gefinancierd door het Onderzoeksfonds
van de Hogeschool Gent (HOF) en het programma Strategisch BasisOnderzoek
van het agentschap voor Innovatie door Wetenschap en Technologie (IWT-SBO)

Gent, 2015

Promotoren:
Prof. dr. Véronique Hoste
Prof. dr. Timothy Colleman

To my parents

Abstract

Applications which make use of natural language processing (NLP) are said to benefit more from incorporating a rich model of text meaning than from a basic representation in the form of bag-of-words. This thesis set out to explore the added value of incorporating deep semantic information in two end-user applications that normally rely mostly on superficial and lexical information, viz. readability prediction and aspect-based sentiment analysis. For both applications we apply supervised machine learning techniques and focus on the incorporation of coreference and semantic role information.

To this purpose, we adapted a Dutch coreference resolution system and developed a semantic role labeler for Dutch. We tested the cross-genre robustness of both systems and in a next phase retrained them on a large corpus comprising a variety of text genres.

For the readability prediction task, we first built a general-purpose corpus consisting of a large variety of text genres which was then assessed on readability. Moreover, we proposed an assessment technique which has not previously been used in readability assessment, namely crowdsourcing, and revealed that crowdsourcing is a viable alternative to the more traditional assessment technique of having experts assign labels.

We built the first state-of-the-art classification-based readability prediction system relying on a rich feature space of traditional, lexical, syntactic and shallow semantic features. Furthermore, we enriched this tool by introducing new

features based on coreference resolution and semantic role labeling. We then explored the added value of incorporating this deep semantic information by performing two different rounds of experiments. In the first round these features were manually in- or excluded and in the second round joint optimization experiments were performed using a wrapper-based feature selection system based on genetic algorithms. In both setups, we investigated whether there was a difference in performance when these features were derived from gold standard information compared to when they were automatically generated, which allowed us to assess the true upper bound of incorporating this type of information.

Our results revealed that readability classification definitely benefits from the incorporation of semantic information in the form of coreference and semantic role features. More precisely, we found that the best results for both tasks were achieved after jointly optimizing the hyperparameters and semantic features using genetic algorithms. Contrary to our expectations, we observed that our system achieved its best performance when relying on the automatically predicted deep semantic features. This is an interesting result, as our ultimate goal is to predict readability based exclusively on automatically-derived information sources.

For the aspect-based sentiment analysis task, we developed the first Dutch end-to-end system. We therefore collected a corpus of Dutch restaurant reviews and annotated each review with aspect term expressions and polarity. For the creation of our system, we distinguished three individual subtasks: aspect term extraction, aspect category classification and aspect polarity classification. We then investigated the added value of our two semantic information layers in the second subtask of aspect category classification.

In a first setup, we focussed on investigating the added value of performing coreference resolution prior to classification in order to derive which implicit aspect terms (anaphors) could be linked to which explicit aspect terms (antecedents). In these experiments, we explored how the performance of a baseline classifier relying on lexical information alone would benefit from additional semantic information in the form of lexical-semantic and semantic role features. We hypothesized that if coreference resolution was performed prior to classification, more of this semantic information could be derived, i.e. for the implicit aspect terms, which would result in a better performance. In this respect, we optimized our classifier using a wrapper-based approach for feature selection and we compared a setting where we relied on gold-standard anaphor–antecedent pairs to a setting where these had been predicted.

Our results revealed a very moderate performance gain and underlined that incorporating coreference information only proves useful when integrating gold-standard coreference annotations. When coreference relations were derived automatically, this led to an overall decrease in performance because of semantic mismatches. When comparing the semantic role to the lexical-semantic features, it seemed that especially the latter features allow for a better performance.

In a second setup, we investigated how to resolve implicit aspect terms. We compared a setting where gold-standard coreference resolution was used for this purpose to a setting where the implicit aspects were derived from a simple subjectivity heuristic. Our results revealed that using this heuristic results in a better coverage and performance, which means that, overall, it was difficult to find an added value in resolving coreference first.

Does deep semantic information help tip the scales on performance? For Dutch readability prediction, we found that it does, when integrated in a state-of-the-art classifier. By using such information for Dutch aspect-based sentiment analysis, we found that this approach adds weight to the scales, but cannot make them tip.

Samenvatting

In toepassingen waar automatische tekstverwerking centraal staat is het beter om te vertrekken van een rijk tekstbegrip dan te vertrouwen op een beperkte lexicale tekstrepresentatie. In dit proefschrift werd de validiteit van deze bewering nagegaan in twee verschillende toepassingen, automatische leesbaarheidspredictie en sentimentanalyse van kenmerken. Voor beide toepassingen werd in het verleden vooral gebruik gemaakt van oppervlakkige en lexicale kennis om voorspellingen te doen. Wij onderzoeken of, en in welke mate, lerende algoritmes tot betere modellen kunnen komen door het toevoegen van diepere semantische kennis, meer bepaald coreferentiële relaties en semantische rollen.

Om dit te kunnen onderzoeken was er in de eerste plaats nood aan systemen voor automatische coreferentieresolutie en voor het automatisch aanduiden van semantische rollen in Nederlandse tekst. De performantie van beide systemen werd geëvalueerd op verschillende tekstgenres. In een volgende stap werden de systemen hertraind op een groot en divers corpus, om tot robuuste modellen te komen voor evaluatie in de twee toepassingen.

Voor de taak van leesbaarheidspredictie werd eerst een tekstcorpus samengesteld dat kon dienen als referentiedataset. We hebben ons hierbij niet beperkt tot een bepaalde tekstsoort, maar selecteerden teksten uit verschillende genres en lieten die beoordelen op leesbaarheid. Daarvoor hanteerden we een voor leesbaarheidsonderzoek nieuwe techniek, crowdsourcen, die uitgaat van het principe dat iedereen met internettoegang een mogelijke annotator is. We hebben kunnen aantonen dat leesbaarheid door zowel taalexperten én leken op dezelfde manier

wordt gelabeld.

We hebben de eerste state-of-the-art automatische leesbaarheidsvoorspeller ontwikkeld. Deze toepassing classificeert twee Nederlandse teksten op basis van hun leesbaarheid en maakt daarbij gebruik van een model waarin zowel lexicale, syntactische als semantische kenmerken van een tekst in overweging worden genomen. In dit doctoraatsonderzoek werd dit achterliggende model nog verder uitgebreid met meer semantische kennis, in de vorm van coreferentiële relaties en semantische rollen. Hun mogelijke meerwaarde werd onderzocht in twee experimenten: eerst werden deze kenmerken manueel toegevoegd of verwijderd, daarna werden de modellen geoptimaliseerd door middel van featureselectie en hyperparameteroptimalisatie. Om deze tweede reeks experimenten te operationaliseren maakten we gebruik van genetische algoritmes. Bij beide experimenten hebben we de resultaten vergeleken wanneer we manueel geannoteerde (*gold standard*) of automatische voorspelde informatie gebruiken over coreferentie en semantische rollen.

Uit onze resultaten blijkt dat de leesbaarheid van teksten wel degelijk beter geclassificeerd wordt wanneer deze diepere semantische kenmerken ook in het model zitten. De beste resultaten werden behaald in het tweede experiment waarin de modellen geoptimaliseerd werden met genetische algoritmes. Tegen onze verwachtingen in haalden de modellen met automatisch voorspelde semantische kennis de beste resultaten. Dit is een interessante bevinding omdat het aantoont dat het mogelijk is om een volledig automatisch leesbaarheidssysteem te ontwikkelen waarin zelfs diepere semantische kennis wordt geïncorporeerd.

In het kader van dit doctoraatsonderzoek werd het eerste end-to-end systeem ontwikkeld voor sentimentanalyse van kenmerken in Nederlandstalige teksten. We verzamelden hiervoor een corpus van online restaurantreviews die werden geannoteerd met informatie rond polariteit en termen die kenmerken (aspecten van het restaurant) in de beoordeling aanduiden. De ontwikkeling van ons systeem bestond uit drie stappen: extractie van termen die kenmerken aanduiden, classificatie van deze kenmerken en classificatie van polariteit. We gingen na of de twee bestudeerde soorten semantische informatie bijdragen tot een betere performantie bij de tweede stap, de classificatie van kenmerken.

In een eerste experiment onderzochten we of coreferentieresolutie kan bijdragen tot correctere classificatie, door te detecteren aan welke expliciete aanduidingen van kenmerken (antecedenten) impliciete aanduidingen (anaforen) gelinkt zijn. We gingen na in hoeverre een baseline classificatiesysteem dat enkel gebruikmaakt van lexicale informatie verbeterd kan worden met extra semantische informatie, namelijk lexicaal-semantische kenmerken en semantische rollen. We vertrokken van de hypothese dat coreferentieresolutie de classificatietaak zou verbeteren omdat het verbanden tussen impliciete en expliciete kenmerken

blootlegt, en er daardoor meer onderliggende semantische informatie kan worden gevonden. We optimaliseerden opnieuw met featureselectie en vergeleken de performantie bij manuele en automatische coreferentieresolutie.

Uit de resultaten blijkt dat het systeem iets beter presteert met deze extra semantische informatie. Hierbij moet wel opgemerkt worden dat coreferentieresolutie enkel bijdraagt wanneer *gold standard* annotaties van coreferentie werden gebruikt. Coreferentie-informatie die automatisch werd gegenereerd deed de performantie van het systeem dalen door fout voorspelde verbanden. Verder werd ook duidelijk dat lexicaal-semantische kenmerken meer bijdragen tot de performantie dan semantische rollen.

In een tweede experiment werd nagegaan hoe impliciete aanduidingen van kenmerken kunnen worden geïdentificeerd. We vergeleken een aanpak waarbij *gold standard* annotaties van coreferentiële relaties werden gebruikt met een aanpak waarbij impliciete kenmerken werden afgeleid met behulp van een eenvoudige subjectiviteitsheuristiek. Hieruit bleek dat deze heuristiek tot betere dekking en hogere performantie van het systeem leidt, en dat coreferentieresolutie hier dus geen meerwaarde biedt.

Kan diepe semantische kennis bijdragen tot betere performantie? Uit ons onderzoek blijkt dat dit het geval is voor leesbaarheidspredictie van Nederlandstalige teksten met een state-of-the-art classificatiesysteem. Bij sentimentanalyse van kenmerken daarentegen resulteert de integratie van semantische informatie niet in een duidelijke stijging van de performantie.

Acknowledgements

At last, the fun part. I am very happy that the intense process of writing a dissertation can culminate in such a small chapter allowing me to thank you all.

I want to start by thanking my supervisor, Véronique Hoste. You have been a true role model and mentor, I owe very much to you and am very grateful for the many chances you have given me. Your ability to always see the bigger picture and your constant belief in me have really helped me across many hurdles, some of which I even considered impossible to take. I also had the pleasure of getting to know you as an open, extremely generous and warm person. Thank you! I look forward to the future and am sure we will share many more wonderful moments.

I am also very much indebted to the other members of my Doctoral Guidance Committee, my co-promotor Timothy Colleman and Lieve Macken. Timothy, thank you for your critical insights and especially for your fresh perspective. Your meticulous proofreading of this PhD was very much appreciated and your comments and suggestions have very much contributed to the end-product. Lieve, it is safe to say that you were the person who persuaded me to go into research. I am thankful for the time you invested in this and other projects we worked on together. Your sound and practical advice were often an eye-opener and your work ethics and resilience (both on a professional and personal level) a true inspiration. I am also very grateful to Antal van den Bosch, Sien Moens and Simone Paolo Ponzetto, who were kind enough to accept the invitation to be in my jury.

Over the years I have had the opportunity to work together with wonderful colleagues on various projects. Willy, Lieve, Piet, Hans and Maribel, thank you very much for the productive DPC cooperation and for allowing me the freedom to take my first steps as a researcher. During the SoNar project I enjoyed the many discussions and talks I had with Martin and of course I am also very much indebted to Paola. I will always remember this project as the tipping point for my career as a researcher, since it opened the doors to the intriguing field of computational linguistics. On the HENDI project I was given the chance to continue the work which was so wonderfully started by Dries and Philip and during the PARIS project I have experienced first-hand how productive and inspiring collaboration amongst researchers from different fields can be. Last year, I spent one semester in Mannheim at the DWS group, thank you Simone for believing in me and insisting that 'I should make an experience out of it', which it most certainly was. Also during many conferences and workshops I have met some wonderful, inspiring people and I would like to thank everyone for the fun moments we shared both on and off these events.

This brings me closer to home. I can truly say that working in Ghent, the city I love so much, and only a few metres from where I grew up (I kid you not), has been amazing. I am very grateful to all the wonderful colleagues at the department that make our buildings such an agreeable and stimulating place to work. A very special thank you is reserved for the fantastic people that are part of the LT3 dreamteam. Els, thank you for brightening up every room with your laughter and for the many confidence boosts, Klaartje thank you for sharing so many thoughts about life and for occasionally even sharing a bed. Marjan and Cynthia, thank you for the very productive and above all very fun collaborations. Peter, Joke, Nils, Mariya, Arda and Julie, thanks for sharing so many laughs during lunch breaks, quidditch or mindfulness sessions to name only a few. I also have very nice memories of some former colleagues. Kathelijne thank you for the many wonderful talks, Sarah for all the help with the normalization pipeline and for being my Pilates buddy. Dries, thank you for staying in touch and for being so brave as to follow your heart.

Then there are these two colleagues that fit somewhere in between the paragraph colleagues and friends. Bart, sharing an office is sharing a life and, man, did we share a lot. Your programming madness, analytical insights, talent to make the most difficult things look easy and especially your humour have coloured many, if not all, of my research days. Isabelle, you and I, woman, we just click. I am so thankful for all your support, for sharing laughs, skincare advice, exciting times and tears (virtual as well as face-to-face). Most of all thank you for inspiring others, including myself, to become a better version of themselves. It is no understatement to say that you both have been my true rocks in the past months and years. Thank you!

The key to life is balance and I am thankful that my dearest friends always made sure that I did not reach my tipping point. Aurore, Inge, Jan, Jon, Jozefien, Kim, Lenz, Lieven, Maaike, Michiel, Mieke, Nele, Pieter, Seba, Tim and Ward. Thank you for sharing growing pains, for being there when the going got rough, for all the joy and laughter and especially for just being who you are. There are two people that I would like to thank more specifically. Evelien, thank you for being my oldest and dearest friend and for always keeping me with both feet on the ground. Thank you for the many discussions we had about what it is being human and how sometimes we don't even need words to understand each other. Sara, my sweetest friend, you have no idea how happy I am that we met on the steps of the Blandijn and how, since that moment, you have become one of the most important parts in my life. Thank you for always sharing so much and for counting on my support for the biggest adventure in life yet to come. I am very excited.

I also want to thank my family. Bart, Veerle, Elisa, Pascal and Mo, thank you for being such great 'in-laws' and for that little sparkle in your eyes. It is wonderful to know and feel that you are all so proud of me. Yvan and Marcella, I have always considered you my second family and I am very grateful for the many wonderful years we could share together, especially in France. Seppe, the true brains of the family, thank you for all the help and for being a wonderful big brother. Lotte, the true workaholic of the family, thank you for making me so proud of everything you already have and still will achieve. Laura and Pieter, thank you for making my siblings, and in return me, so happy and loved. And then there are my parents, who have been there all along. Thank you for giving us so many opportunities, for putting us first, for our carefree childhood and for always being there, always. Above all, thank you for teaching us the most important lesson in life: to live it to the fullest. This book is dedicated to you both.

My final words are for Koen, who else. Thank you for all the love, for giving me so much freedom and for sharing so much silliness and joy. Know that everything would be trivial if it were not for you. Thank you!

Contents

CHAPTER 1

Introduction

Applications that make use of natural language processing (NLP) benefit more from a rich model of text meaning rather than from a basic representation in the form of bag-of-words and n-gram models (Hovy et al. 2006, Jurafsky and Martin 2008). The most well-known and famous example is probably IBM's Watson, a deep question answering system that was able to beat humans on the Jeopardy quiz by integrating both shallow and deep knowledge (Ferrucci et al. 2010).

This dissertation focuses on exploring the added value of incorporating deep semantic information in a readability prediction system and an aspect-based sentiment analysis pipeline.

In reading research the focus has long been on developing formulas relying on superficial text characteristics. These formulas remain very popular, and form a substantial, and often the only, part of every new readability prediction system introduced in the market. But what exactly determines the readability of a given text? Can we rely solely on characteristics such as word or sentence length to determine a text's readability? Or should we also try to measure other more intricate characteristics such as syntactic complexity or coherence?

In aspect-based sentiment analysis, all sentiment expressions within a given document and the concepts and aspects to which they refer have to be detected, making it a very fine-grained sentiment analysis task.

Current state-of-the-art systems rely mostly on bag-of-words information for detecting these aspects. But are words enough? Can more information about the entities and their roles expressed in a text help to pinpoint the different agents and aspects?

We try to answer these questions by incorporating semantic and discourse information, in the form of semantic roles and coreference, into both a readability prediction and an aspect-based sentiment analysis system. Semantic roles specify the roles of entities in a particular text and allow us to abstract from the specific lexical expressions denoting these. They can be derived at the clause level and can represent various semantic aspects of the relation between a predicate and its arguments. Coreference on the other hand tells us something more about which entities refer to the same referent in a text and hence form a coreference chain. Encoding coreference helps NLP systems to look beyond the level of single sentences and pay attention to discourse, an ability which is also considered crucial for successful end-user systems such as question answering or automatic summarization (Webber and Joshi 2012).

1.1 Background

The effectiveness of superficial lexical features such as bag-of-words and token and character ngram models has proven difficult to overrule using more complex linguistic knowledge. In the field of information retrieval, for example, there used to be an established consensus that little can be gained from complex linguistic processing for tasks such as text categorization and search (Moschitti and Basili 2004). At the same time, however, a considerable amount of research in the field of NLP, has been devoted to developing such deep linguistic resources and systems.

Our focus is on coreference resolution and semantic role labeling. For both tasks, much advances were made by the organization of shared tasks and challenges such as the MUC-6 (1995), MUC-7 (Chinchor 1998) or the more recent SemEval-2010 (Recasens et al. 2010) and CoNLL (Pradhan et al. 2012) shared tasks devoted to coreference resolution and the CoNLL 2004 and 2005 (Carreras and Màrquez 2004, 2005) or SemEval 2007 (Baker et al. 2007) shared tasks devoted to semantic role labeling. Besides these challenges, large corpus projects emerged, such as the OntoNotes corpus where approximately one million words have been annotated with syntactic and semantic structures and coreference (Hovy et al. 2006). Though English remains the most-resourced language, research on other, smaller languages, lagged not far behind. For Dutch, a noteworthy initiative in this respect was the creation of the SoNaR 1 corpus in the framework of the STEVIN-programme (Spyns and Odijk 2013). This corpus comprises one million words and thus presents the first Dutch corpus integrating multiple levels of annotation, including coreference and semantic roles.

Nevertheless, the added value of incorporating these two deep processing techniques in end-user application remains understudied (Poesio et al. 2010, Màrquez et al. 2008). In this dissertation, we focus on exploring the added value of coreference and semantic roles in two such applications, viz. readability prediction and aspect-based sentiment analysis.

Readability research and the automatic prediction of readability has a long tradition. Whereas superficial text characteristics leading to on-the-spot readability formulas were popular until the last decade of the previous century (Flesch 1948, Gunning 1952, Kincaid et al. 1975), recent advances in the field of computer science and natural language processing have triggered the inclusion of more intricate characteristics in present-day readability research (Si and Callan 2001, Schwarm and Ostendorf 2005, Collins-Thompson and Callan 2005, Heilman et al. 2008, Feng et al. 2010). When it comes to current state-of-the art readability prediction systems, it can be observed that even though more complex features trained on various levels of complexity have proven quite successful when implemented in a readability prediction system (Pitler and Nenkova 2008, Kate et al. 2010, Feng et al. 2010), there is still no consensus on which features are actually the best predictors of readability. As a consequence, when institutions, companies or researchers from other disciplines wish to use readability prediction techniques, they still rely on the more outdated superficial characteristics and formulas (van Boom 2014).

The domain of sentiment analysis is a relatively new strand of NLP research, concerned with modeling subjective information in text. The field has seen rapid expansion in recent years and its focus has shifted from coarse-grained opinion mining on the document-level to fine-grained sentiment analysis, where the sentiment is assigned at the clause level (Wilson et al. 2009). In this respect, aspect-based sentiment analysis (Pontiki et al. 2014) focuses on the detection of all sentiment expressions within a given document and the concepts and aspects (or features) to which they refer. Such systems do not only try to distinguish positive from negative utterances, but also strive to detect the target of the opinion, which comes down to a very fine-grained sentiment analysis task. State-of-the-art systems tackling this task rely almost exclusively on lexical features (Pontiki et al. 2014). Moreover, though the potential added value of coreference resolution is pointed out in many survey works (Liu 2012, Feldman 2013), qualitative research on the added value of actually incorporating this kind of information is scarce.

1.2 Motivation

From a theoretical computational linguistic point of view it is important to find out how and to what extent semantic roles and cohesion in the form of coreference can be modeled in text. We can ask ourselves the question: how are these two deep semantic layers actually realized in text and what kind of linguistic and extralinguistic (world) knowledge is required to model these?

From an application point of view, the question remains whether it is beneficial to incorporate these deep semantic processing steps into end-user applications. With respect to our two tasks, we can formulate the questions whether the coherence of a text can be assessed using coreference resolution and whether semantic roles contribute to predicting the readability of a given text. Or, when we resolve anaphor–antecedent pairs in consumer reviews, does this allow us to derive more aspects, and can more (semantic) information on an aspect help to pinpoint the different agents and aspects present in a review?

Finally, if this deep semantic processing information does seem to contribute to better overall performances, the question remains whether the current state-of-the-art is sufficient to incorporate these processing steps in the pipeline? In other words, what level of accuracy should these semantic processing systems be able to attain?

1.3 Research objectives

In accordance with the research aims described above, the main research question of this study can be formulated as follows:

Can deep semantic processing in the form of coreference resolution and semantic role labeling lead to better models for automatic readability prediction and aspect-based sentiment analysis?

This research question consists of three large buildings blocks, presented next, each raising a more specific research question that needs to be answered in the framework of this dissertation.

1.3.1 Deep semantic processing

Before we can start implementing deep semantic information in the form of coreference and semantic role information in end-user applications, we first need systems capable of deriving this kind of information. To this purpose, we adapt a Dutch state-of-the-art coreference resolution system, COREA, and develop a semantic role labeler for Dutch, SSRL. For both systems we use a supervised machine learning approach and rely on memory-based learning.

Most state-of-the-art coreference resolvers and semantic role labelers are trained and tested on one genre, namely newspaper text. We, however, want to build a system capable of predicting the readability of texts that language users are generally confronted with on on a more or less daily basis, ranging from newspaper articles to mortgage files, i.e. very diverse text material. In addition, for the aspect-based sentiment analysis, we will work with customer reviews, which means we will be confronted with a very specific text genre, namely user-generated content.

It is thus of key importance that our underlying semantic processing techniques are robust enough. Many tools reveal a drop in performance when tested on data belonging to a different genre than the one the system was trained on (Daumé III et al. 2010). This is why we test the cross-genre portability of the COREA and SSRL system using a large corpus of semantically annotated data comprising a variety of genres, viz. the Dutch SoNaR one-million-word corpus (SoNaR 1), leading to our first more specific research question:

RQ 1: How robust are coreference resolution and semantic role labeling systems when applied to a large variety of text genres?

1.3.2 Readability prediction

As mentioned above, we wish to implement a generic readability prediction system capable of assessing a large variety of text material. In this respect, we build the first classification-based system for Dutch. For the construction of such a system using a supervised machine learning approach, three steps can be roughly distinguished. First of all, a readability corpus containing text material of which the readability will be assessed must be composed. Second, a methodology to acquire readability assessments has to be defined. Finally, based on the readability corpus and the acquired assessments, prediction tasks can be performed.

The investigation of the readability of a wide variety of texts without targeting a specific audience, has not received much attention (Benjamin 2012). There exist almost no general domain corpora, especially not for Dutch, and other methodologies, apart from having experts assign labels, are scarce. We compile the first general evaluation corpus of Dutch generic text comprising various text genres and levels of readability. Moreover, we propose and apply a completely new assessment technique which has not yet been used in readability assessment, namely crowdsourcing, and compare this technique to the use of expert labels.

We build the first classification-based readability prediction system for Dutch. We distinguish a binary and more fine-grained multiclass classification task and incorporate a range of state-of-the-art information sources (or features) that have proven useful to predict readability. We want to push the state of the art by incorporating coreference and semantic role information. We distinguish automatically-predicted and gold-standard coreference and semantic role features in order to discover the true upper bound of adding this kind of deep semantic information to a readability prediction system.

We evaluate model performance and investigate which semantic information sources are appropriate for both classification tasks by manually including or excluding these features. We seek to optimally exploit the discriminative power of the semantic features and explore a wrapper-based approach to feature selection using genetic algorithms, something which has not been investigated in readability research before.

This can be translated to our second research question:

RQ 2: Can we push the state of the art in generic readability prediction by incorporating deep semantic text characteristics in the form of coreference and semantic role features?

1.3.3 Aspect-based sentiment analysis

Aspect-based sentiment analysis has proven important for mining and summarizing opinions from online reviews. Several English benchmark datasets have been made publicly available. For Dutch, however, no such benchmark datasets exist. We collect a corpus of restaurant reviews and annotate this corpus by adapting the guidelines developed for one of those benchmark datasets to Dutch.

We develop the first aspect-based sentiment analysis pipeline for Dutch by distinguishing three individual subtasks: aspect term extraction, aspect category classification and aspect polarity classification. This means that, first, aspect terms are automatically derived, in a next step they are assigned to a correct aspect category, and finally their polarity is classified. We implement and filter the output of an existing end-to-end terminology extraction system for the first subtask and develop multiclass classifiers for the second and third subtask.

By incorporating semantic information sources, in the form of lexical-semantic and semantic role features, we build a more complex model for aspect category classification than simply relying on lexical information. We evaluate model performance and seek to optimally exploit the discriminative power of these semantic features by applying a wrapper-based approach to feature selection using genetic algorithms. In this respect we also compare a setting where coreference resolution was performed prior to classification to one where it was not. We distinguish gold-standard and automatically-derived coreference relations in order to assess the true upper bound of including this type of information.

We also test the fully-automatic pipeline and are the first to perform a qualitative analysis of whether resolving coreference results in an added value for the task of aspect-based sentiment analysis. We investigate this by first deriving explicit aspect terms, after which implicit aspect terms are derived using coreference information and, we compare this to using a simple heuristic to this purpose. In a final step, both the explicit and implicit aspect terms are classified into aspect categories after which their polarity is assigned.

Our third more specific research question can be defined as:

RQ 3: Does more information on discourse entities and their roles help to pinpoint the different aspects and aspects in aspect-based sentiment mining?

1.4 Thesis outline

This thesis consists of eleven chapters and is structured as follows. Chapter 2 discusses the two semantic information layers that form the starting point for the research described in this dissertation. It explains which systems were adapted and developed to assign these two sources of information and how the cross-genre robustness of these systems was tested on a large corpus comprising a variety of text material.

The remainder of the dissertation can be divided into two large parts. Part I explores how our two deep semantic processing techniques can be implemented as additional features in a readability prediction system and how their added value can be tested. Part II has a similar objective but focuses on the field of aspect-based sentiment analysis. In this respect, only semantic roles can be implemented as additional features, whereas coreference resolution can be used to resolve implicit aspects. Both parts consist of four individual chapters.

Part I: Readability Prediction

Chapter 3 provides an introduction to the field of readability research and discusses existing work on supervised machine learning approaches to readability prediction, with a special focus on the features that have been investigated in previous research.

Chapter 4 describes the corpus of Dutch general texts that was collected for this study. A large part of this chapter is devoted to the exploration of a new technique for assessing readability. We compare a traditional readability assessment technique, i.e. consulting expert readers, to the use of crowdsourcing.

In Chapter 5, we discuss the information sources or features that were implemented. We built a state-of-the-art readability prediction system including both superficial, lexical, syntactic and semantic features. Our main focus is on these semantic features. In the next part of this chapter, we discuss the two classification tasks that were performed: a binary task in which the readability of two text was compared and a more fine-grained task where more subtle differences in readability had to be classified. We introduce the models that we developed and describe how the added value of our deep semantic processing techniques was validated using joint optimization.

All experimental results are presented and discussed in Chapter 6. We establish to what extent coreference and semantic role features contribute to our two classification tasks, and what the impact is of the various optimization strategies. We perform a qualitative error analysis and end this chapter with a conclusion of the first part of this dissertation.

Part II: Aspect-based Sentiment Analysis

Chapter 7 introduces the field of opinion mining or sentiment analysis and discusses existing work, with a special focus on the task of aspect-based sentiment analysis and on how the output of semantic role labeling and coreference resolution systems has been implemented in previous research.

Chapter 8 presents the corpus that was collected comprising Dutch restaurant reviews, which were all annotated following established guidelines. Aspect-based sentiment was annotated on a sentence-per-sentence basis by first indicating individual aspect terms and grouping these into predefined aspect categories, after which the sentiment expressed towards these aspects was annotated.

Chapter 9 presents the pipeline of the first aspect-based sentiment analysis system that was developed for Dutch texts consisting of three large incremental subtasks: aspect term extraction, aspect term category classification and polarity classification. We discuss how we tackled each of these individual steps and devote specific attention to the incorporation of our two deep semantic processing layers. Next, we introduce the models that were developed to assess the added value of these two deep semantic information layers.

All experimental results are presented and discussed in Chapter 10. We explain and extensively discuss to what extent processing the data with coreference resolution prior to classification and incorporating semantic roles as additional semantic features can help for the subtasks of aspect term extraction and aspect category classification. We perform an error analysis on each individual subtask and finish this chapter with a conclusion of this second part.

Chapter 11 presents the overall conclusions of this thesis and outlines perspectives for future work.

CHAPTER 2

Deep semantic processing

In this chapter, we introduce the two semantic layers, the possible added value of which will be closely investigated for Readability Prediction (Part I) and Aspect-Based Sentiment Analysis (Part II). The mainstream paradigm in computational semantics today is to let the computer automatically learn from corpora, i.e. machine learning (Koller and Pinkal 2012). We explain the two systems that were adapted and retrained following this paradigm in Section 2.1 and Section 2.2. In Section 2.3 we test the robustness of these tools on a large semantically-annotated corpus comprising various text genres. Section 2.4 concludes this chapter.

2.1 Layer one: coreference

Coreference is a pervasive phenomenon in natural language and is one of the fundamental ingredients of semantic interpretation (Poesio et al. 2010). Coreference resolution is the task of automatically recognizing which words or expressions refer to the same discourse entity in a particular text or dialogue. By building coreference chains, we can identify all relevant information about all the entities present in a text. Consider the following description on Wikipedia[1]:

(1) [Barack Hussein Obama II] (born August 4, 1961) is [the 44th and current President of the United States], and [the first African American to hold the office]. Born in Honolulu, Hawaii, [Obama] is [a graduate of Columbia University and Harvard Law School], where [he] served as [president of the Harvard Law Review]. [He] was a [community organizer in Chicago] before earning [his] law degree. [He] taught constitutional law at the University of Chicago Law School from 1992 to 2004. [He] served three terms representing the 13th District in the Illinois Senate from 1997 to 2004, running unsuccessfully for the United States House of Representatives in 2000.

Thanks to coreference we can infer a lot about the entity 'Barack Hussein Obama II'. In this example all the items in between square brackets refer to the same entity, and together they form a coreference chain.

Intuitively, the identification of coreference links seems crucial for applications and it has indeed proven a useful resource for automatic summarization (Steinberger et al. 2007), information extraction (Hendrickx et al. 2013) and opinion mining (Jakob and Gurevych 2010). However, in other studies the added value of coreference or anaphora resolution is less clear, e.g. for question answering (Morton 2000) and summarization, textual entailment and text classification (Mitkov et al. 2012).

There exists an immense body of work dedicated to the task of coreference resolution as described in surveys by Mitkov (2002), Strube (2009), Poesio et al. (2010) and Ng (2010). With the appearance of publicly available coreference corpora as part of the MUC-6 (1995) and MUC-7 (Chinchor 1998) conferences, machine learning techniques for coreference resolution became popular in the 1990s. The initial focus of the MUC and other shared tasks (e.g. ACE (Doddington et al. 2004) and the i2b2 challenge[2]), was on English text and two genres in particular: the news and medical genres. But soon after that, other languages and genres were explored, too (ACE 2005, SemEval-2010 (Recasens et al. 2010) and CoNLL (Pradhan et al. 2012)). Next to the benchmark datasets produced for these challenges, many language-specific corpora and treebanks emerged as stand-alone projects in the beginning of the new millennium, such as the Tübingen (Telljohann et al. 2004) and Prague Dependency (Hajič et al. 2006) treebanks for German and Czech or the NAIST (Iida et al. 2003) and AnCora-CO (Recasens and Martí 2010) corpora for Japanese and Spanish/Catalan.

[1] Retrieved from http://en.wikipedia.org/wiki/Barack_Obama [03-30-2015].

[2] https://www.i2b2.org/NLP/Coreference/Call.php

The same applies to Dutch. Two corpora annotated with coreferential relations were developed in the first decade of the 21st Century: the KNACK-2002 corpus (Hoste and De Pauw 2006), comprising only newspapers texts and the COREA corpus (Hendrickx et al. 2008), comprising both KNACK-2002 and other news texts, speech transcripts and medical texts. For Dutch, the annotation of different genres culminated in the SoNaR 1 corpus (which will be presented in Section 2.3.1). Besides SoNaR, a smaller corpus containing Dutch user-generated content (blogs and news comments) was annotated with coreference too (Hendrickx and Hoste 2009).

The most widespread machine learning approach to coreference resolution is the mention-pair model. Basically, this is a binary classifier that decides for every two NPs or mentions within a text whether they are coreferent or not and performs clustering afterwards[3]. Though first proposed by Aone and Bennett (1995) and McCarthy and Lehnert (1995), its breakthrough came with the successful approach of Soon et al. (2001) which was further developed by Ng and Cardie (2002). Their choice of features, training and decoding methods became the standard benchmarking baseline for coreference resolution (Poesio et al. 2010). Other models have been proposed, such as the entity-mention model which tries to determine whether an NP or mention is coreferent with a preceding, possibly partially formed cluster (Luo et al. 2004, Yang et al. 2008) or the ranking model that determines which entity is the most likely antecedent given an NP to be resolved (Connolly et al. 1994, Denis and Baldridge 2008, Rahman and Ng 2009). In addition, unsupervised systems have been developed or combinations of supervised, unsupervised and rule-based systems (Haghighi and Klein 2009, Lee et al. 2013). The most successful of these is Stanford's multi-pass sieve coreference resolution system which performs entity-centric coreference. In this system all mentions that point to the same real-world entity are jointly modeled, in a rich feature space using solely simple, deterministic rules (Lee et al. 2013).

For English, it is not easy to pinpoint the current state of the art in coreference resolution since it is difficult to compare the many systems that have been developed, all using their own corpora and scoring mechanisms (Ng 2010). The MUC datasets are seen as a benchmark against which most of the systems working with English have been compared. On the MUC-6 test set the best results starting from automatically predicted mentions are reported by supervised systems, a MUC F-score of 68.4 by Ng and Cardie (2002) and one of 71.3 by Yang et al. (2003). For more details on the evaluation of coreference, we refer to Section 2.3.2.

In the next section we describe the coreference resolver for Dutch named COREA (Hoste 2005, Hendrickx et al. 2008), which also implements a mention-pair

[3] See Section 2.1.1 for further details.

model. We explain how the original COREA system works and present the adaptations we have made in the framework of this dissertation. It is this tool the cross-genre robustness of which we test later on and the output of which we incorporate in both readability prediction and sentiment analysis experiments. The best results using the original system were reported by Hoste (2005), viz. a MUC F-score of 51.4 on the KNACK-2002 corpus, which comprises only newspaper text.

2.1.1 The COREA coreference resolver

We restrict the task of coreference resolution to the resolution of identity relations between nominal constituents, e.g. noun phrases, including names, and pronouns (NPs). An identity relation implies that two NPs refer to the same discourse entity. Identity relations can be distinguished from three other relations, namely bridge, bound and predicative relations. For more information on the different coreference relations we refer to Section 2.3.1.

If we consider the following example:

(2) NL: Onder de Vlaamse Primitieven is [Jan van Eyck] (Maaseik, ca. 1390 - Brugge, 1441) ongetwijfeld de voornaamste meester. [Hij] is een Vlaams kunstschilder.
EN: Amongst the Flemish Primitives, [Jan van Eyck] (Maaseik, ca. 1390 - Bruges, 1441) is by far the most pronounced master. [He] is a Flemish painter.

Our tool will be able to predict the identity relation that exists between the two NPs indicated in between square brackets 'Jan van Eyck' and 'Hij', referring to the same discourse entity.

In COREA, coreference resolution is seen as a classification task in which each pair of NPs in a text is classified as having a coreferential relation or not. For each pair, a feature vector is created representing the characteristics of that particular pair and the relations between the NPs.

Though the original system extracted NPs based on chunk information, we adapted it so that it can identify NPs based on the output of a richer information source, i.e. dependency parse trees using Alpino (van Noord et al. 2013). This enables a more fine-grained recognition of NPs in that nominal constituents are extracted based on a deep grammatical parsing instead of extracting noun phrase recognized by a shallow chunker[4]. For our running example, our system is able to extract the following nine NPs:

(de Vlaamse Primitieven, Jan van Eyck, Maaseik, 1390, 1441, Brugge, de voornaamste meester, Hij, een Vlaams kunstschilder)

The dependency output also includes the output of various preprocessing steps: tokenization, lemmatization and Part-of-Speech (PoS) tagging. Besides these information sources, named entity recognition is also performed using MBT (Daelemans et al. 2003), trained on the CoNNL shared task Dutch dataset (Tjong Kim Sang 2002) and an additional gazetteer lookup.

With all this information feature vectors are created for every possible NP pair capturing the information about the two NPs under consideration and the relations between these two. The pairs itself are made by linking every NP to its preceding NP, with an upper limit of going 20 sentences backwards. We will now briefly describe and illustrate the features that would be derived for the pair 'Hij' and 'Jan van Eyck'. For a more detailed description we refer to Hoste (2005).

- the **distance** between the noun phrases is expressed in the number of sentences, number of intervening NPs and a binary feature encoding whether this NP distance is larger than two.
  (1 5 1)
- the **local context** of the anaphor expressed by the three lemmata before and after the anaphor and their corresponding part-of-speech tags. If the anaphor occurs at the beginning or end of a sentence, these features are represented with the symbol '=='.
  (== == == == == == is een Vlaams WW(pv,tgw,ev) LID(onbep,stan,agr) ADJ(prenom,basis,zonder))
- **syntactic** information coding whether the two NPs are in an apposition relation, and a feature representing their syntactic function. An additional binary feature indicates whether the syntactic function of the two NPs is the same.
  (0 subject subject 1)
- **morphosyntactic** information encoding several properties of the anaphor, the antecedent and their relation. In total, fourteen features indicate whether they are pronouns, proper nouns, demonstrative or reflexive pronouns, and, if applicable, the number of the pronoun, and whether they

[4] This technique was already introduced for facilitating the annotation process in the COREA project itself (Bouma et al. 2007), but it had not been implemented in the system before.

are definite or indefinite nouns.
(1 0 0 0 0 0 0 1 0 1 3p 0 1 0)

- **matching** features describing whether both NPs have the same gender, number and whether there is string overlap in the form of an alias (e.g. United States – US) and a complete, partial or head string overlap.
(na num_na 0 0 0 0)
- **semantic** features are expressed in the form of named entity information, synonym and hypernym lookup in Cornetto (a Dutch database combining Dutch WordNet (Vossen 1998) and the Referentie Bestand Nederlands (Martin and Ploeger 1999)) and semantic cluster overlap (based on clusters extracted with unsupervised k-means clustering on the Twente Nieuws Corpus by Van de Cruys (2005)).
(person person 1 0 0 0 0 0)
- **class** if we have training data available, this feature indicates whether the NPs are coreferent or not.
(POS)

For the actual classification, we make use of the TiMBL algorithm (Daelemans et al. 2009) since this was the learner achieving the best results on the KNACK-2002 dataset, i.e. a MUC F-score of 51.4 (Hoste 2005).

TiMBL is a memory-based learning algorithm, using the *k* Nearest Neighbor method which stores all examples in memory during training. At classification time, a previously unseen text example is presented to the model which then looks for the *k* most similar examples – nearest neighbors - in memory and performs an average of their classes in order to predict a class label. TiMBL's value of the *k* value differs in that it refers to the k-nearest distances instead of k-nearest examples. This is done because several examples in memory can be equally similar to a new instance. In this way, instead of choosing one at random, all equal examples at the same distance are added to the set of nearest neighbors (Daelemans and van den Bosch 2005).

When instances are classified in an unseen test document, multiple NPs might be labeled as positive, when actually only one should be.

Consider the following example:

(3) NL: Naast [Jan van Eyck] wordt [Rogier Van der Weyden] beschouwd als de belangrijkste schilder van de 15e eeuw. [Hij] was wellicht de invloedrijkste schilder van die eeuw. Van der Weyden voegde emotie toe aan de Vlaamse schilderkunst.

> EN: Rogier Van de Weyden is considered, together with Jan van Eyck, the most important painter of the 15th century. He was probably the most influential painter of that century. Van der Weyden added emotion to Flemish painting.

Here the NP pairs 'he – Rogier Van der Weyden' and 'he – Jan van Eyck' might both be classified as positive instances. To solve this problem, the task of coreference resolution continues after classification. In a next step, coreference chains need to be built for the NP pairs that were classified as coreferential. Instead of selecting one single antecedent per anaphor, the COREA system builds complete coreference chains for a document based on overlap. In this respect it differs from previous approaches such as the 'closest first' (Soon et al. 2001) or 'most likely' approach (Ng and Cardie 2002).

In order to create the complete coreference chains we use the counting mechanism as proposed in Hoste (2005):

1. Given an instancebase with anaphor - antecedent pairs (ana$_i$, ant$_{ij}$), for which i = 2 to N and $j = i$ 1 to 0. Select all positive instances for each anaphoric NP. Then make groupings by adding the positive ant$_{ij}$ to the group of ana$_i$ and by adding ana$_i$ to the group of ant$_{ij}$. The following is an example of such a grouping. The numbers represent IDs of anaphors/antecedents. The number before the colon is the ID of the anaphor/antecedent and the other numbers represent the IDs which relate to this anaphor/antecedent.

   2: 2 5 6 25 29 36 81 92 99 231 258 259 286

   5: 2 5 6 25 29 36 81 92 99 231 258 259 286

   6: 2 5 6 25 29 36 81 92 99 231 236 258 259 286

   8: 8 43 64 102 103 123 139 144 211 286

   20: 20 32 69 79

2. Then compare each ID grouping with the other ID groupings by looking for overlap between two groupings and select the pairs with an overlap value above a predefined threshold.

   2: **2 5 6 25 29 36 81 92 99 231 258 259 286**

   5: **2 5 6 25 29 36 81 92 99 231 258 259 286**

   6: **2 5 6 25 29 36 81 92 99 231** 236 **258 259 286**

   8: 8 43 64 102 103 123 139 144 211 **286**

   20: 20 32 69 79

If we compute the overlap between the grouping of ID 2 with the groupings of IDs 5, 8 and 20, for example, we observe a complete overlap of groupings 2 and 5. Combining ID 8 with ID 2, however, leads to a very weak overlap (only on one ID) and an overlap value of 0.08. No overlap is found for the combination of ID 20 and ID 2. If we take into account an overlap threshold of 0.1, this implies that the two last NP pairs will not be selected.

3. For each pair with an overlap value above the threshold, we compute the union of these pairs which results in an incremental construction of coreference chains.

For more details, we refer to Hoste (2005). We made no adaptations to this coreference chain construction step.

2.2 Layer two: semantic roles

The analysis of semantic roles within a text is concerned with the characterization of events, such as determining *who did what to whom, when, where and how*. Semantic roles are indicated at the clause level and the first step is to find predicate-argument structures. This is not a trivial task, because a lot of variation can exist in the syntactic realizations of semantic arguments. A semantic role is actually the theoretical concept relating syntactic complements and semantic arguments (Koller and Pinkal 2012). The predicate of a clause (typically a verb) establishes *what* took place, and other sentence constituents express the participants in the event (such as *who*), as well as further event properties (such as *when*, *where* and *how*).

The primary task of semantic role labeling (SRL) is to indicate exactly which semantic relations exist between a predicate and its associated participants and properties (Màrquez et al. 2008).

Let us consider the following example sentence:

(4) He taught constitutional law at the University of Chicago Law School from 1992 to 2004.

Here, we observe one predicate ('taught') describing what takes place. One participant can be distinguished 'He' and some additional event properties are described relating to location and time.

Until now, in linguistics, there is no agreement on a definitive list of semantic roles or even on the question whether it would be possible to compile such an exhaustive list at all. Some major semantic roles are agreed on, such as the Agent (which would be the 'He' in our example) or Theme. Many lists have been proposed, such as the situation-specific roles suggested by Fillmore et al. (2004), the thematic set of general roles as proposed by Jackendoff (1990) or even a set of only two core roles, a Proto-Agent and Proto-Theme (Dowty 1991). These linguistic approaches have also influenced the computational work on SRL, leading to the creation of computational lexicons capturing the foundational properties of predicate–argument relations. The two most well-known are FrameNet (Fillmore et al. 2003) and PropBank (Palmer et al. 2005) which both triggered a substantial body of work.

Semantic role labeling has proven beneficial for NLP applications such as information extraction (Surdeanu et al. 2003). It is also beneficial for automatic summarization (Melli et al. 2005) and machine translation (Liu and Gildea 2010, Gao and Vogel 2011). Moreover, it has shown to increase the number of questions that can be handled in question answering systems (Narayanan and Harabagiu 2004, Shen and Lapata 2007) and to improve textual entailment in that it enables complex inferences that are not allowed using surface representations (de Salvo Braz et al. 2005, Sammons et al. 2009). The use of SRL systems in such real-world applications, however, has been rather limited (Màrquez et al. 2008), which makes it an interesting task to research.

When it comes to semantic interpretation research, the rise of statistical machine learning methods in NLP in the 1990s also invigorated research in this field. This started with the automatic learning of subcategorization frames (Briscoe and Carroll 1997) or classifying verbs according to argument structure properties (Merlo and Stevenson 2001, Schulte im Walde 2006). But as soon as medium-to-large corpora were manually annotated with semantic roles such as FrameNet (Fillmore et al. 2003), PropBank (Palmer et al. 2005) or NomBank (Meyers et al. 2004) research on automatic semantic role labeling (SRL) really took off. The first statistical machine learning approach to SRL was developed by Gildea and Jurafsky (2002), trained on FrameNet. This study initiated much similar research, but in the following years PropBank came to replace FrameNet as the most popular resource because it provides a more representative sample of text annotated with semantic roles, i.e. the Penn Treebank (Marcus et al. 1993), compared to the manually selected examples as presented in FrameNet. A lot of progress for English was made with the organization of shared tasks such as the CoNLL 2004 and 2005 Shared tasks (Carreras and Màrquez 2004, 2005), centering around PropBank and the Senseval-3 (Litkowski 2004) and SemEval 2007 shared tasks (Baker et al. 2007), considering frame semantic parsing.

Ever since the seminal work of Gildea and Jurafsky (2002), semantic role labeling has been perceived as a task in which two steps are performed: argument identification and argument classification. Previous research has shown that for the first step syntactic knowledge is important whereas the second one necessitates more semantic information (Pradhan et al. 2008). For the first step of argument identification one thus has to decide which basic syntactic representation to follow. For Dutch, Monachesi et al. (2007) were among the first to choose dependency over constituent syntax because of its rich syntactic information and ability to provide very useful information on the relation between parts of a sentence such as grammatical functions.[5] After both the CoNLL 2008 (Clark and Toutanova 2008) and 2009 (Hajič 2009) tasks were devoted to this subject, using dependency structures now seems to have become common practice. For the second step, argument classification, the most common approach is to build a classifier and describe the information between a predicate and its arguments using various features. The CoNLL 2004 and 2005 shared tasks were based on PropBank and represent the most frequently used evaluation benchmark for English, on which the best systems obtained an F1 score of ca. 80%. See Section 2.3.3 for more information on the scoring of semantic role labeling.

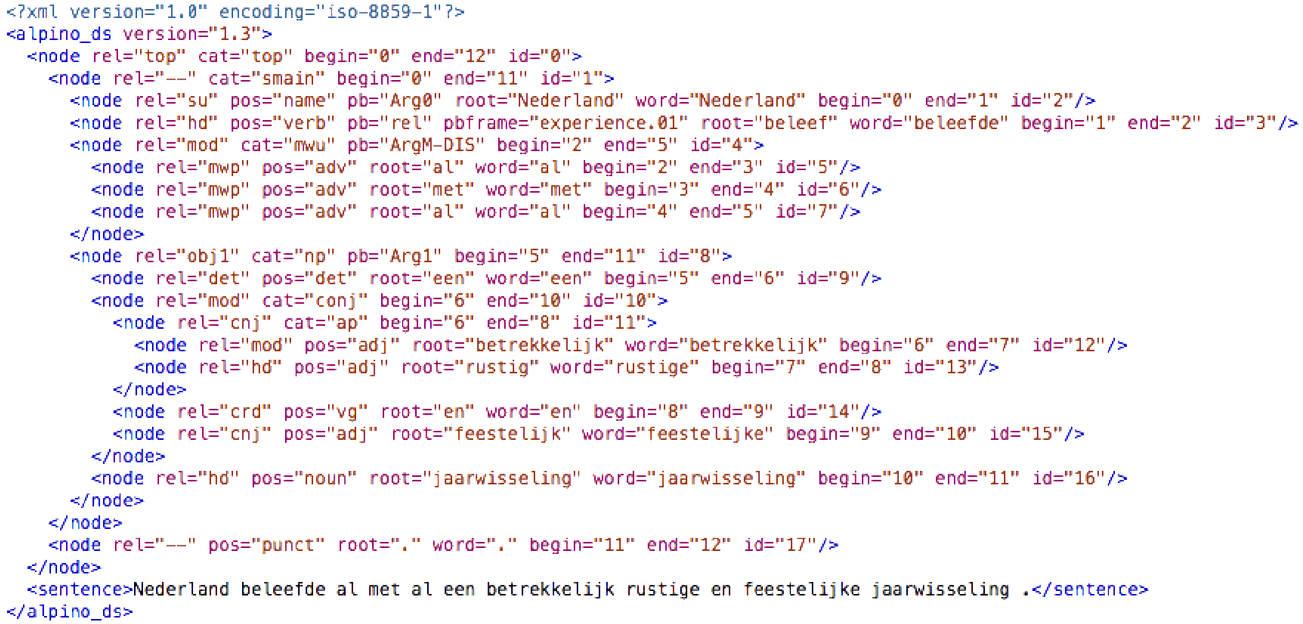
For Dutch, a first semantic role labeler was developed by Stevens et al. (2007) and a very small set of example sentences were annotated in the framework of the D-Coi project (Trapman and Monachesi 2006), but it was not until the appearance of the SoNaR 1 corpus (cfr. infra) that we retrained this tool on a substantial amount of training data and further improved it. This tool, the SoNaR semantic role labeler (SSRL) which also follows the PropBank approach is presented in the next section.

2.2.1 The SoNaR semantic role labeler

Following the seminal approach by Gildea and Jurafsky (2002) we treat semantic role labeling as a two-step task consisting of argument identification and classification. For the identification step, the system relies on the output of Alpino (van Noord et al. 2013), which generates dependency structures. In a first step, these dependency structures are used to detect the predicates within each sentence. Next, we link every predicate to its possible arguments. It should be noted that we only consider verbs as predicates and only siblings of the verbs within the dependency structure can be considered as arguments.

An example of the dependency structures were are dealing with is presented in Figure 2.1.

[5] For a full discussion we refer to Johansson and Nugues (2008)



```
<?xml version="1.0" encoding="iso-8859-1"?>
<alpino_ds version="1.3">
  <node rel="top" cat="top" begin="0" end="12" id="0">
    <node rel="--" cat="smain" begin="0" end="11" id="1">
      <node rel="su" pos="name" pb="Arg0" root="Nederland" word="Nederland" begin="0" end="1" id="2"/>
      <node rel="hd" pos="verb" pb="rel" pbframe="experience.01" root="beleef" word="beleefde" begin="1" end="2" id="3"/>
      <node rel="mod" cat="mwu" pb="ArgM-DIS" begin="2" end="5" id="4">
        <node rel="mwp" pos="adv" root="al" word="al" begin="2" end="3" id="5"/>
        <node rel="mwp" pos="adv" root="met" word="met" begin="3" end="4" id="6"/>
        <node rel="mwp" pos="adv" root="al" word="al" begin="4" end="5" id="7"/>
      </node>
      <node rel="obj1" cat="np" pb="Arg1" begin="5" end="11" id="8">
        <node rel="det" pos="det" root="een" word="een" begin="5" end="6" id="9"/>
        <node rel="mod" cat="conj" begin="6" end="10" id="10">
          <node rel="cnj" cat="ap" begin="6" end="8" id="11">
            <node rel="mod" pos="adj" root="betrekkelijk" word="betrekkelijk" begin="6" end="7" id="12"/>
            <node rel="hd" pos="adj" root="rustig" word="rustige" begin="7" end="8" id="13"/>
          </node>
          <node rel="crd" pos="vg" root="en" word="en" begin="8" end="9" id="14"/>
          <node rel="cnj" pos="adj" root="feestelijk" word="feestelijke" begin="9" end="10" id="15"/>
        </node>
        <node rel="hd" pos="noun" root="jaarwisseling" word="jaarwisseling" begin="10" end="11" id="16"/>
      </node>
    </node>
    <node rel="--" pos="punct" root="." word="." begin="11" end="12" id="17"/>
  </node>
  <sentence>Nederland beleefde al met al een betrekkelijk rustige en feestelijke jaarwisseling .</sentence>
</alpino_ds>
```

Figure 2.1: Example of an Alpino dependency structure containing semantic roles (pb–attributes).
EN: The Netherlands experienced overall a considerate quiet and festive new year.

In this dependency tree, the predicate and semantic roles are indicated with the 'pb-attribute'. As one can notice, the predicate (*rel*) and the semantic roles (*Arg0*, *ArgM-DIS* and *Arg1*) all occur on the same level in the dependency tree, i.e. they are siblings.

This identification step leads to a large number of predicate-argument pairs that are the input for our classifier. For the actual classification we extract a number of features that describe properties of the predicate, argument and the relation between these two and indicate the semantic role (if there exists one).

Let us consider the following example:

(5) NL: Alleen de gewone man betaalt belastingen.
EN: Only average Joe pays taxes.

Based on the Alpino output our system is able to extract two predicate-argument pairs from this sentence:

(betaalt-Alleen de gewone man) and (betaalt-belastingen).

As was the case with our COREA system, we are also able to derive the output of various preprocessing steps from Alpino, i.e. tokenization, lemmatization and part-of-speech tagging. All this information allows us to extract various predicate and argument features.

- **predicate** features:
  - predicate lemma: the predicate/verb's lemma.
  - predicate PoS-tag: In order to reduce the CGN-tagset (Van Eynde 2005), which originally consists of 320 distinct tags, we only used a word's particular main class and one subclass. E.g. V(pv, enk, zijd) = V(pv)
  - predicate voice: binary feature to code the voice of the predicate (active/passive).

- **argument** features:
  - argument c-label: the category label of the possible argument (NP, PP, ...)
  - argument d-label: the dependency label (e.g. Mod,...)
  - argument position: a binary feature indicating whether the argument is positioned before or after the predicate.

- argument head word: the lemma of the argument's head word, if the argument is no leaf node this is looked up based on the dependeny labels
- argument head word PoS-tag: the PoS-tag of the head word, again we opted for a less fine-grained labeling. E.g. N(soort,ev,basis,zijd) = N(soort).
- argument's first word + PoS-tag and argument's last word + PoS-tag: if an argument consists of more than one word we looked up the first and last lemma of the word and their corresponding PoS-tags.
- CAT/POS pattern: the left-to-right chain of d-labels of the argument and its siblings
- REL pattern: the left-to-right chain of c-labels of the argument and its siblings
- CAT + REL pattern: the c-label of the argument concatenated with its d-label

This would result in the following two feature vectors for our running example:

betaalt – Alleen de gewone man
betalen,ww(pv),active,np,su,before,man,n(soort),alleen,BW(),man,N(soort), np*verb*noun,su*hd*obj1,su*np,Arg0

betaalt – belastingen
betalen,ww(pv),active,#,obj1,after,belastingen,n(soort),#,#,#,#,np*verb*noun, su*hd*obj1,obj1*,Arg1

The last feature is the label indicating which semantic role the instance represents, if there is one. Considering the possible semantic roles we follow the PropBank approach and thus distinguish between four arguments and ten modifiers. We refer to the next section for more information on the possible labels. The classification task is thus one of multiclass classification. For our SSRL system, we used TiMBL as our default machine learning algorithm to perform this task. Since this is the first time that the SSRL has been trained on such a large dataset we do not have top-performing scores for this tool yet.

2.3 Cross-genre robustness experiments

The focus of this dissertation is in exploring the added value of incorporating deep semantic information in two end-user applications which currently rely mainly on superficial text characteristics. We use the two systems that were introduced in the previous sections in order to answer two of our central research questions:

1. Can we push the state of the art in generic readability prediction by incorporating deep semantic text characteristics in the form of coreference and semantic role features?
2. Does more information on discourse entities and their roles help to pinpoint the different aspects and aspects in aspect-based sentiment mining?

For the Readability Prediction experiments (Part I) we envisage to build a system capable of predicting texts we are all confronted with on an average day, ranging from newspaper articles to mortgage files, i.e. very diverse text material. In the second part we investigate Aspect-Based Sentiment Analysis (Part II) of customer reviews, which means we will be confronted with a very specific text genre, namely user-generated content.

One of the challenges in many NLP tasks is to test their portability across different genres. This is important because many tools, especially those using a supervised machine learning paradigm reveal a drop in performance when tested on data belonging to a different genre than the one the system was trained on (Daumé III et al. 2010). Most current coreference resolvers and semantic role labelers are also trained and tested on one genre, namely newspaper text (for example the MUC (1995) and ACE (2004) datasets for English and the KNACK-2002 corpus for Dutch coreference resolution or the Penn Treebank (Marcus et al. 1993) comprising Wall Street Journal texts which was annotated with semantic roles in PropBank (Palmer et al. 2005)).

As our two envisaged applications will typically work with non-newspaper text material, it is of key importance that the underlying semantic processing techniques are robust enough. This is why we decided to test the cross-genre portability of COREA and SSRL using a large corpus of semantically annotated data comprising a variety of genres, viz. the Dutch SoNaR one-million-word corpus (SoNaR 1). We start by introducing this corpus in Section 2.3.1. Next, we describe the experiments that were conducted to evaluate the cross-genre portability of our two systems under consideration: COREA (Section 2.3.2) and SSRL (Section 2.3.3). This research has been published in De Clercq et al. (2011, 2012)

2.3.1 Introducing the SoNaR 1 corpus

The lack of an effective digital language infrastructure for Dutch was the starting point of the STEVIN-programme which funded research projects that should allow researchers in linguistics and computational linguistics to perform corpus-based research (Spyns and Odijk 2013). One of those STEVIN-funded projects was the SoNaR project in which a large reference corpus of contemporary written Dutch has been built comprising a wide variety of traditional text genres and texts coming from new media (Oostdijk et al. 2013).

An important part of SoNaR is the one-million-word subcorpus, SoNaR 1. This core corpus had already been enriched with manually verified part-of-speech tags, lemmatization and syntactic analysis in previous research (van Noord 2009) and during the SoNaR project four additional semantic layers were added and manually verified: named entities, coreferential relations, semantic roles and spatio-temporal relations. This corpus thus presents the first Dutch corpus integrating multiple levels of annotation and it is this corpus that we used to test the cross-genre robustness of our coreference resolver and semantic role labeler.

We start by giving some more details on the genre subdivision within this corpus after which we explain how coreference and semantic roles have been annotated in the framework of the SoNaR-project.

Genres

For the genre subdivision within SoNaR 1 we rely on the definition of genre as formulated by Biber (1988, p. 170):

> Genre categories are determined on the basis of external criteria relating to the speaker's purpose and topic; they are assigned on the basis of use rather than on the basis of form.

These external criteria can be things such as the intended audience, purpose and activity type. In other words, it refers to conventional, culturally recognized groupings of texts based on properties other than lexical or grammatical (co-)occurrence features (Lee 2001).

Within SoNaR 1 we discern the following six genres:

- The **administrative** genre which consists of reports, speeches and minutes of meetings which are all intended for internal communication within companies or institutions;

- The **autocues** genre containing written newswire intended primarily for the hearing impaired;
- The genre referred to as **external communication** represents website material, press releases and newsletters, all intended for a broad external audience;
- The **instructive** genre includes manuals, patient information leaflets and procedure descriptions which are intended for a broad audience but mostly consist of more technical information;
- The **journalistic** genre consists mainly of newspaper articles which are intended to inform the general public about current affairs.
- The sixth genre has data originating from Dutch **wikipedia**, which also aims to inform the broader audience but has a more encyclopedic purpose.

These six genres form the basis for the cross-genre experiments we performed, which will be explained in the following sections.

Coreference annotation

For the coreferential relation annotations, the complete corpus of one million words was manually annotated following the guidelines developed by Hoste and De Pauw (2006) and Bouma et al. (2007). These guidelines allow for the annotation of four relations and special cases are flagged. The four annotated relations are identity (NPs referring to the same discourse entity), bound (expressing properties of general categories), bridge (as in part-whole, superset-subset relations) and predicative.

If we consider the example,

(6) NL: Onder de Vlaamse Primitieven is Jan van Eyck (Maaseik, ca. 1390 - Brugge, 1441) ongetwijfeld de voornaamste meester, een vernieuwer op het gebied van de landschaps- en portretschildering. Hij is een Vlaams kunstschilder.
EN: Amongst the Flemish Primitives, Jan van Eyck is by far the most pronounced master, a forerunner in landscape and portrait painting. He is a Flemish painter.

In this example three predicative (Jan van Eyck – de voornaamste meester, Jan van Eyck – een vernieuwer op het gebied van de landschaps- en portretschildering and Hij – een Vlaams kunstschilder) and one identity relation (Jan van Eyck – Hij) would be annotated.

Moreover, the guidelines allow for the flagging of special cases: negations and expressions of modality, time-dependency and identity of sense (as in the so-called paycheck pronouns (Karttunen 1976)[6]). As annotation environment, the MMAX2 annotation software[7] was used. Some annotation statistics are presented in Table 2.1.

Type	**Number**
Documents	861
Tokens	1,000,437
Identity relations	81,547
Predicative relations	34,213
Bound relations	197
Bridge relations	3,568

Table 2.1: Annotation statistics indicating the amount of data that was annotated and how many instances of each relation were found.

Semantic roles annotation

Contrary to coreference annotation, for the semantic roles only a very small subset had been annotated in the past, i.e. 3000 predicates were annotated during the D-Coi project[8] (Schuurman et al. 2010). This is why it was decided during the SoNaR-project to only manually verify 500,000 words. All data were annotated with semantic roles following the PropBank approach (Palmer et al. 2005). The PropBank guidelines were adapted to Dutch by Trapman and Monachesi (2006).

The semantic roles are indicated at clause level and a distinction is made between a predicate and its arguments. The predicate is the semantic head of a sentence. Within SoNaR 1 only verbs were considered as predicates and labeled as such. Instead of creating new Dutch framefiles, everything was mapped

[6] Named after the following example: 'The man who gave [his paycheck]$_1$ to his wife was wiser than the man who gave [it]$_1$ to his mistress.' Clearly *his paycheck* and *it* do not refer to the same object in the world which means there is no identity relation between both NPs, but there is a semantic overlap.

[7] http://mmax2.net

[8] Actually, the D-Coi project was a pilot project which was inititiated to study the feasibility of creating a Dutch reference corpus.

onto the English PropBank frameset[9]. Next, the elements related by these specific predicates, the arguments (Arg), were annotated. In addition, modifiers (ArgM), which add additional semantic information, were labeled as well.

If we consider the example:

(7) NL: Alleen de gewone man betaalt belastingen.
EN: Only average Joe pays taxes.

This example would be annotated as:
Alleen de gewone man (Arg0) betaalt (PRED = pay.01)[10] belastingen (Arg1).

As annotation environment TrEd[11] was used. Some annotation statistics are presented in Table 2.6, which at the same time constitutes an exhaustive list of all semantic role labels distinguished following the PropBank approach. The additional information added to every possible label was derived from the PropBank guidelines (Babko-Malaya 2005).

All this gold-standard annotated data could be used to retrain the two systems the output of which we wish to use for our experiments later on: the COREA system which is able to perform noun-phrase coreference resolution of identity relations and the SSRL which is able to enrich dependency trees with semantic roles. We will explain how we tested these two systems' cross-genre robustness in order to find out what type and amount of training data is most beneficial for the robustness of both systems.

2.3.2 Testing COREA's robustness

Experimental setup

Thanks to the SoNaR 1 corpus we have access to over one million words annotated with coreference spread over six different genres. In other projects, medical (Hendrickx et al. 2008) texts, consisting of entries from a medical encyclopedia, and user-generated content, in the form of weblogs and comments to news articles (Hendrickx and Hoste 2009) have also been annotated.

This is why we decided to compare COREA's performance on these eight different text genres: administrative (ADM), autocues (AUTO), external communication (EXT), instructive (INST), journalistic (JOUR), medical (MED), user-

[9] https://verbs.colorado.edu/propbank/framesets-english/
[10] As found on http://verbs.colorado.edu/propbank/framesets-english/pay-v.html
[11] http://ufal.mff.cuni.cz/∼pajas/tred/

Type	**Number**	*Additional information*
Documents	354	
Sentences	29,432	
Tokens	500,850	
Predicates	36,979	*Semantic head*
Arg0	18,323	*external argument (proto-agent)*
Arg1	31,537	*internal argument (proto-patient)*
Arg2	7,079	*indirect object, beneficiary, instrument, attribute, end state*
Arg3	506	*start point, beneficiary, instrument, attribute*
Arg4	601	*end point*
ArgM-ADV	5,155	*Adverbials*
ArgM-CAU	1,590	*Causal clauses*
ArgM-DIR	556	*Directionals*
ArgM-DIS	5,342	*Discourse markers*
ArgM-EXT	922	*Extent markers*
ArgM-LOC	6,777	*Locatives*
ArgM-MNR	4,904	*Manner markers*
ArgM-MOD	5,391	*Modals*
ArgM-NEG	2,947	*Negation*
ArgM-PNC	1,803	*Purpose markers*
ArgM-PRD	1,164	*Secondary predicates*
ArgM-REC	1,205	*Reciprocals*
ArgM-TMP	10,097	*Temporals*

Table 2.2: Annotation statistics indicating the amount of data that was annotated and how many predicates, arguments and modifiers were found in the data plus some more explanation regarding the different PropBank labels.

generated content (UGC) and wikipedia (WIKI) articles. To rule out data size as a possible explanation for performance shifts, datasets of equal size (about 30K tokens[12]) were randomly selected. As explained in Section 2.1 our system only considers identity relations. Table 2.3 gives some statistics about each dataset, such as the average sentence length and the number of coreferring NPs.

All datasets were preprocessed following the COREA pipeline. Tokenisation, lemmatisation, part-of-speech tagging and grammatical relations were derived from the manually verified output of the Alpino parser, i.e. gold-standard dependency structures[13]. Only for the UGC data no gold-standard dependency

[12] This number was chosen because the MED genre only contained 30,001 tokens

[13] In a real-world setting we will not have this gold standard information available, but for these experiments it allows us to isolate the performance of our coreference resolution system

	# Documents	# Tokens	avgsentencelen	#Coreferring
ADM	21	30,215	18.1	2,403
AUTO	15	30,058	14.6	2,411
EXT	29	29,940	15.9	2,381
INST	18	29,994	17.5	3,024
JOUR	52	30,002	18.2	2,472
MED	213	30,001	14.4	1,995
UGC	56	29,740	19.7	3,063
WIKI	15	30,340	18.9	3,480

Table 2.3: Size and number of coreferring NPs per genre dataset.

trees were available which necessitated us to use automatically labeled dependency structures with Alpino. The best reported result of the Alpino parser is a mean accuracy of 89.17 when trained and tested on standard text material (van Noord et al. 2013). Named entity recognition was performed using MBT (Daelemans et al. 2003), trained on the 2002 CoNNL shared task Dutch dataset (Tjong Kim Sang 2002) and an additional gazetteer lookup.

As features, the system encodes distance, local context, syntactic, lexical and semantic information (see Section 2.1.1). We build instances between all NP pairs going 20 sentences back in context. NPs that are not part of a coreferential chain (*singletons*) are thus included as negative examples but not as chains of consisting of only one element. For all experiments we used TiMBL version 6.3 (Daelemans et al. 2010) with default parameter settings.

An important consideration when assessing performance is to ensure that the validation is carried out on a representative test set. We opted for cross-validation. This is a generalization of data splitting where p instances are omitted from the data, and their classes are predicted with a model trained on the remaining n - p instances. The process is repeated until all n instances have been tested. This approach allows all the data to be used for both training and testing. We employ k-fold cross validation, where the data is randomly split into k subsets of roughly the same size, called folds. A model is trained on k - *1* folds and tested on the remaining fold, and the process is rotated k times (Weiss and Kulikowski, 1991). Since 10-fold cross validation is the de facto standard (Jurafsky and Martin 2009), for our experiments k was set to 10.

The results are evaluated using the four scoring metrics as implemented in the scoring script from the coreference resolution task of the SemEval-2010 competition (Recasens et al. 2010).

by ruling out mistakes made in preprocessing stages.

- The MUC scoring software (Vilain et al. 1995) counts the number of links between the coreferential elements in the text, and looks how many links are shared or not between the gold-standard coreferential chains and the system predictions. As MUC concentrates on links, elements that are not part of a coreferential chain, but that are mentioned only once (*singletons*), are not taken into account in this scoring method.

- The B-cubed measure (Bagga and Baldwin 1998) does not consider mere links between elements, but takes into account the coreferential clusters of elements referring to the same entity. B-cubed computes for every individual element in the text the precision and recall by counting how many elements are in the true coreferential cluster and how many in the predicted coreferential cluster.

- The CEAF measure (Luo and Zitouni 2005) focuses on a one-to-one mapping of elements in the true and predicted coreferential clusters. Both B-cubed and CEAF measures are sensitive to the presence of many singletons: the larger the percentage of singletons, the higher these scores (Recasens and Hovy 2011).

- More recently, the BLANC measure (Recasens and Hovy 2011) was developed to overcome problems with the other scoring methods. This measure is a variant of the Rand Index (Rand 1971) adapted for coreference resolution and it averages over a score for correctly detecting singletons, and a score for detecting the correct cluster for coreferential elements.

It should be noted that our system follows the basic MUC criterion that two entities are only linked when they are coreferential, so it does not take into account chains of only one element (singletons). As a consequence, contrary to the SemEval-2010 competition, when we compute our scoring metrics, a singleton that is erroneously classified as part of a coreference chain is counted as an error. When a mention is correctly indicated as a singleton, however, this is not represented in the scores because our system does not label these in the final step. This should be held in mind when interpreting the results.

In order to test cross-genre robustness, we ran three sets of experiments:

1. In the first set of experiments, we wanted to investigate whether adding more data is beneficial for the classifier. We trained the classifier on each genre individually and compared performance with different training set sizes. Three experiments were conducted: we first trained on each individual genre and tested on the relevant genre using ten-fold cross validation (each fold 27K vs. 3K). In a second experiment, the classifier was trained

on all genres except one and tested on the one that was left out (210K vs. 30K). In a third experiment, we used all data, including genre-specific training material for training the classifier, in a ten-fold cross validation setup (each fold 237K vs. 3K).

2. In a second set of experiments, we focused on the actual cross-genre portability. In order to test this, we each time trained on a single genre and tested the performance of the classifier trained on this single genre on each of the other genres.
3. Based on the results obtained in the second batch of experiments, we investigated in this set of experiments whether the inclusion of particular genres when training on all data actually decreases performance. In other words, does excluding outlier genres from training data increase performance? This was done by each time leaving out the genre exhibiting the worst cross-genre portability and performing ten-fold cross validation.

Results

The results of the **first set** of experiments are presented in Figure 2.2. The dots marked as *individual* present the experiments in which each classifier was trained and tested on the same material. The scores for *All-individual* present experiments in which the classifiers are trained on a large and diverse training set of all different genres except the genre that is held out as a test set. The last experiments in the graph *All+individual* show the result when training on all genres including the held-out genre.

From these graphs we can conclude that MUC, B-Cubed and CEAF scores present the same tendency, even though the B-cubed and CEAF scores are lower than MUC (B-cubed and CEAF range from 20 to 35, whereas MUC ranges from 25 to 60).

These three graphs reveal that adding more diverse training material improves performance (*All-Individual* is almost always better than *Individual*, except for the MED genre). This improvement is even more outspoken when a small part of genre-specific information is also included to the full training set: the best results are always achieved in the *All+Individual* setup, this time also for the MED genre.

BLANC, however, seems to contradict the other metrics. Though the scores are higher (they range in between 45 and 60), they reveal that according to BLANC, a larger training set containing both genre-specific and different data (*All+Individual*, which was the best setup according to the other three metrics)

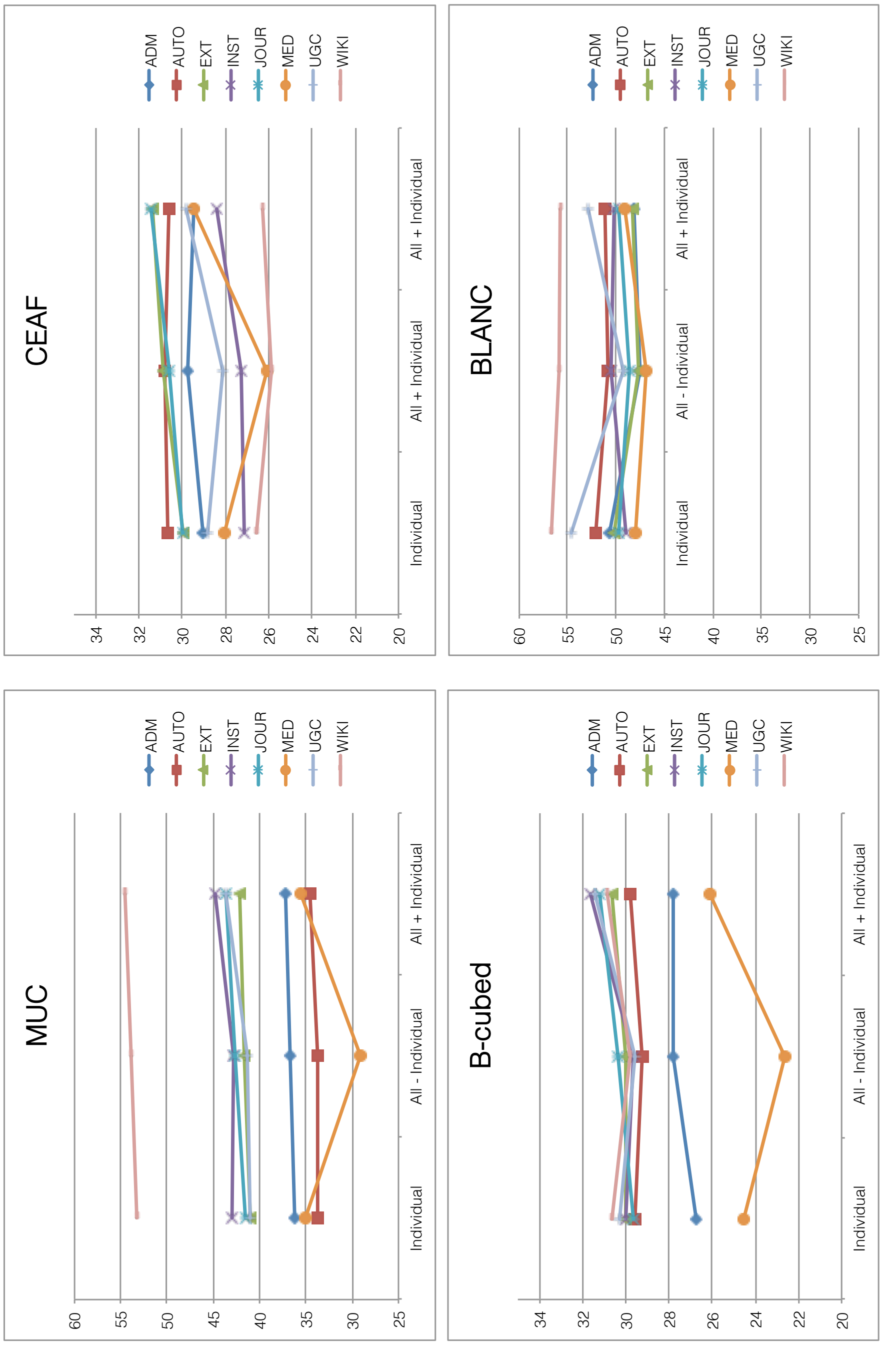


Figure 2.2: Performance comparison for each genre when training only on that genre (Individual), all the other genres except that one genre (All - Individual), or both (All + Individual) respectively.

is not the optimal setting. BLANC suggests that training on genre-specific material only is the best approach (i.e. in five cases of the eight genres).

This brings us to the **second set** of cross-genre experiments, where we each time train on one genre and test on all the other genres individually until all genres have once been used as training data.[14] In order to represent the results, we ranked the classifier performance on each genre, ranging from the genre-classifier which on average performs worst when being applied to the other genres, to the one performing best. We performed this ranking for each of the four evaluation metrics.

The final ranking is visualized in Table 2.4 below.

MUC	B3	CEAF	BLANC
MED	MED	MED	MED
UGC	UGC	UGC	UGC
INST	INST	INST	INST
EXT	EXT	EXT	JOUR
WIKI	AUTO	AUTO	ADM
AUTO	ADM	ADM	AUTO
ADM	WIKI	WIKI	EXT
JOUR	JOUR	JOUR	WIKI

Table 2.4: Comparison of the worst (top) to best-performing (bottom) genres per metric.

Although there are some differences between the metrics – we again observe that BLANC tends to differ more from the others in its ranking – they all seem to agree that MED (medical text), UGC (unedited text) and INST (instructive text) constitute poor cross-genre training material. JOUR has been selected by MUC, B3 and CEAF as the best material for training on other genres. Overall, the four metrics confirm that three genres have less generalization power, viz. MED, UGC and INST.

In the **third set** of experiments, we aim to optimize our selection of training data to get the best possible overall performance. We hypothesize that leaving out those genres with the least predictive power from the training material will increase overall performance. In this set of experiments we train on all data, including genre-specific information, and test on one genre while progressively leaving out the three above-mentioned genres. The results of this *reversed learning curve* for all metrics can be found in Table 2.5. Whenever a score is printed in bold, it is the best score obtained for a particular genre.

[14] Train on ADM = test on AUTO; train on ADM = test on UGC;....

It is difficult to compare the different metrics with each other. We observe that only the BLANC metric confirms our expectation that the results are almost always better when poor training material (in the form of the MED, UGC and INST material) is excluded from training. The results as measured with the other three metrics, however, reveal that leaving out this type of data is only beneficial for half of the datasets. An important observation to make is that, for all metrics, the performance gains which are obtained by leaving out data are modest. In other words, if we remove poor training data from our instancebase this does not (always) seem to help. As a consequence, these numbers do not strongly confirm our hypothesis.

To conclude, we can state that in order to get a good generalization performance it is more important to have a larger training set comprising a variety of text material than to put additional time and effort in the composition of this training set. In order to explore the added value of incorporating coreference information in a readability prediction system and an aspect-based sentiment analyser we used a version of COREA trained on the 500,000 words annotated in the framework of SoNaR 1 consisting of six different text genres.

2.3.3 Testing SSRL's robustness

Experimental setup

Thanks to SoNaR 1 we have access to over 500,000 words which are manually labeled with semantic roles, representing six genres. Contrary to our coreference resolution setup, no datasets had been labeled with semantic roles before for Dutch. We thus tested the cross-genre robustness of our semantic role labeler on the six genres present within SoNaR: administrative (ADM), autocues (AUTO), external communication (EXT), instructive (INST), journalistic (JOUR) and wikipedia (WIKI).

To rule out data size as a possible explanation for performance shifts, again, datasets of equal size (about 50K) were randomly selected. Table 2.6 gives some statistics about each dataset, such as the number of tokens, sentences and the number of labeled predicates.

	ADM	AUTO	EXT	INST	JOUR	MED	UGC	WIKI
MUC								
ALL	37.10	34.61	42.09	**44.81**	43.63	**35.57**	**43.61**	**54.48**
1MinMED	37.26	34.41	42.01	44.61	44.03		43.56	54.07
2MinUGC	**37.39**	**34.85**	**42.29**	44.51	**44.56**	35.44		54.35
3MinINST	37.06	34.00	41.81		44.46	34.72	31.02	54.21
B-cubed								
ALL	27.83	**29.77**	**30.64**	**31.66**	31.23	**26.08**	31.45	**30.84**
1MinMED	27.74	29.64	30.18	**31.66**	31.34		**31.68**	30.46
2MinUGC	**28.02**	29.46	30.11	31.26	**31.81**	25.99		30.58
3MinINST	27.87	29.54	30.01		31.61	25.18	31.02	30.64
CEAF								
ALL	29.48	**30.61**	**31.36**	28.42	31.42	**29.49**	29.79	26.31
1MinMED	29.11	30.33	30.26	**28.47**	30.86		29.96	**26.40**
2MinUGC	**29.73**	29.51	30.09	28.12	**31.62**	29.33		25.99
3MinINST	29.58	30.48	29.16		30.93	28.20	**29.97**	25.14
BLANC								
ALL	48.10	51.11	48.29	50.21	49.74	**49.01**	52.87	55.73
1MinMED	48.49	51.37	48.51	50.72	49.55		**54.70**	**56.66**
2MinUGC	48.73	51.49	48.73	**51.01**	**50.37**	48.15		56.11
3MinINST	**49.71**	**51.59**	**50.88**		49.61	48.49	54.16	56.17

Table 2.5: Results of the third set of experiments for all metrics and in comparison with training on all data.

Datasets	#Tokens	#Sentences	#Predicates
ADM	50,123	2,591	3,234
AUTO	50,155	3,233	4,100
EXT	50,122	2,584	3,182
INST	50,072	1,654	4,514
JOUR	50,020	2,339	3,645
WIKI	50,000	3,296	2,816

Table 2.6: Data statistics indicating the number of tokens, sentences and labeled predicates present in each dataset.

Following the SSRL pipeline all datasets were preprocessed in the same way. As with the coreference experiments, we could rely on gold-standard dependency trees to derive our candidate experiments. Besides dependency syntax, these dependency trees also provided us with preprocessing information in the form of tokenization, lemmatization and part-of-speech tagging.

For the next step of argument classification, the system encodes as features properties of both the arguments and predicates. For all experiments we used Timbl version 6.3 (Daelemans et al. 2010) with default parameter settings and again performed ten-fold cross-validation experiments.

We evaluate by calculating precision, recall and F-measure. Each time, we present the overall scores for argument classification. Precision measures the ratio of correct assignments by the classifier, divided by the total number of the classifier's assignments (Equation 2.1). It is a measure for the amount of irrelevant hits that a system produces, i.e. how noisy it is. Recall, also called sensitivity, is defined to be the ratio of correct assignments by the classifier divided by the total number of gold-standard assignments (Equation 2.2).

$$precision = \frac{true\ positives}{true\ positives\ +\ false\ positives} \tag{2.1}$$

$$recall = \frac{true\ positives}{true\ positives\ +\ false\ negatives} \tag{2.2}$$

F-score is a measure that combines precision and recall, as per Equation 2.3. The β term determines the weight of precision versus recall. Traditionally, F-score is calculated with $\beta = 1$, resulting in a harmonic mean of precision and recall, known as F_1 score or balanced F-score (Equation 2.4). We report F_1 score for our experiments, since it is the standard and most widespread F-score implementation.

$$F_{\beta} = \left(1+\beta^{2}\right)\cdot\frac{precision\cdot recall}{\beta^{2}\cdot precision\ +\ recall} \tag{2.3}$$

$$F_{\mathit{1}} = 2\cdot\frac{precision\cdot recall}{precision\ +\ recall} \tag{2.4}$$

The main objective of the experiments was to find out what works best for our classifier: training on genre-specific data or on a more diverse dataset incorporating a variety of genres. In order to do so, we conducted three sets of 10-fold cross-validation experiments on the 50K datasets. In order to allow for comparison, the 5K test set partitions were kept constant over all experiments.

1. In a first experiment, the classifier was trained on the 45K genre-specific training partitions and tested on the respective 5K test partitions.
2. In a second experiment, the robustness of the classifier was evaluated by exclusively training the classifier on information coming exclusively from different genres and in a next stage we compare this to a setup where both genre-specific and not genre-specific data are included.
3. In the third set we add more data to the labeler and perform two experiments. In a first experiment, we included data from different genres (5 genres * 50K = 250K) and tested the performance of the classifier on the 5K test partitions mentioned in the previous sets. In a final step, genre-specific data was also added to the training set (5 genres * 50K + 1 genre * 45K = 295K).

Results

The results of the **first** set of experiments are presented in Table 2.7. We notice that training on a small amount of genre-specific data already yields pretty high results, ranging from F_{1} measures of 71.75 for the journalistic to 77.32 for the instructive genre.

Whereas in the first set of experiments, we wanted to illustrate how the labeler would perform when it was trained on a dataset which was specifically tailored to the test set at hand (i.e. training and testing on one and the same genre), the second experiment aimed at completely the opposite.

In the **second** set of experiments, the results of which are presented in Table 2.8, we first investigated how the labeler would perform when trained only on data

Datasets	Prec.	Rec.	F$_1$
ADM	73.21	72.63	72.92
AUTO	72.63	72.19	72.40
EXT	72.12	71.11	71.61
INST	77.53	77.12	77.32
JOUR	72.14	71.36	71.75
WIKI	73.05	71.85	72.44

Table 2.7: Results of training and testing on genre-specific data using 10-fold cross validation.

from other genres than the genre of the test set (*All-Individual*). This should allow us to draw some conclusions regarding the robustness of our semantic role labeler. If we compare these results to the results of the previous experiment (Table 2.7), we observe a drop in performance on all genres, with scores dropping by 2 to 4%.

	All-Individual			*All+Individual*		
Datasets	Prec.	Rec.	F$_1$	Prec.	Rec.	F$_1$
ADM	70.26	70.00	70.13	74.05	73.90	**73.97**
AUTO	70.16	68.28	69.21	73.10	72.19	**72.64**
EXT	70.92	70.18	70.55	73.31	72.48	**72.89**
INST	73.54	73.09	73.31	77.88	77.64	**77.76**
JOUR	71.52	70.65	71.08	72.98	72.35	**72.66**
WIKI	71.01	69.60	70.30	73.65	72.35	**72.99**

Table 2.8: Results of training on different genres and a combination of genre-specific and not genre-specific data using 10-fold cross validation.

This led us to additional experiments in which we also included a small sample, viz. 1/6 of the 50K, of genre-specific training data instead of only focussing on other genres (*All+Individual*). The results for these experiments are presented on the right-hand side of Table 2.8 and reveal an improvement for all genres, even outperforming the genre-specific experiments presented in Table 2.7.

These findings had to be corroborated on larger datasets though, and that is why we performed a **third** set of experiments where more data was added to our labeler. The same experiments as in the previous set were conducted, but with the difference that now all available training data from the other genres was also included into training. The results of these final experiments are presented in Table 2.9

	All-Individual			*All+Individual*		
Datasets	Prec.	Rec.	F$_1$	Prec.	Rec.	F$_1$
ADM	72.48	72.75	72.61	74.67	74.69	**74.68**
AUTO	73.14	72.00	72.57	74.65	73.73	**74.19**
EXT	73.57	73.01	73.29	74.48	73.71	**74.09**
INST	75.61	76.06	75.83	79.18	79.18	**79.18**
JOUR	73.47	72.92	73.19	74.47	73.89	**74.18**
WIKI	74.52	73.71	74.11	75.87	75.03	**75.45**

Table 2.9: Results of training on genre-specific and not genre-specific data using 10-fold cross validation.

The overall best results are achieved with this last setup. We see that when we only include a larger amount of not genre-specific data (*All-Individual*) only three genres benefit from this (the external, instructive and wikipedia genres). If we also include a small amount of genre-specific information, however, we see that for all genres the best results are achieved. This strengthens our belief that including more training data is beneficial but that including genre-specific training information is also necessary. We clearly notice that including only a small amount of genre-specific data accounts for the best results.

As with the coreference resolution experiments we can conclude that in order to get a good generalization performance it is more important to have a larger training set comprising a variety of text material than to put additional time and effort in the composition of this training set. As a consequence, we also retrained SSRL on the 500,000 words annotated in the framework of SoNaR 1, which consists of six different text genres.

2.4 Conclusion

In this chapter, we introduced the two deep semantic information layers – coreference and semantic roles – the added value of which we wish to investigate in two end-user applications. Coreference tells us something more about the relations that can exist between entities in a particular text and semantic roles help us to determine *who did what to whom, when, where and how* on a sentence-per-sentence basis.

We explained the two systems that were adapted and retrained in order to perform coreference resolution and semantic role labeling using a supervised machine learning paradigm. The coreference resolution system, COREA, is able to perform noun-phrase coreference resolution of identity relations. The semantic role labeler, SSRL, enriches dependency trees with semantic roles.

The main part of this chapter was dedicated to investigating the cross-genre portability of these two systems using a large corpus of semantically annotated data comprising a variety of text genres, the SoNaR 1 corpus. This was necessary because we envisaged from the beginning that the two end-user applications we wish to implement – Readability Prediction (Part I) and Aspect-Based Sentiment Analysis (Part II) – will be applied to a large variety of text material.

Based on these cross-genre experiments, we can conclude that for both the coreference resolution and semantic role labeling system, it is advisable to train on a substantial amount of training data comprising various text genres. If possible, it is best to also include a (small) amount of genre-specific training data. This is why we trained both systems on a variety of text material that had been manually annotated with both coreference and semantic roles in the framework of the SoNaR 1 corpus comprising six distinct text genres: administrative texts, autocues, texts used for external communication, instructive texts, journalistic texts and wikipedia texts. The amount of training data was kept constant for our two systems, i.e. 500,000 words.

Part I

Readability Prediction

CHAPTER 3

Preliminaries

In the first two chapters, we discussed that although many advances have been made in NLP and the development of deep linguistic processing techniques, the incorporation of these techniques in end-user applications is not a straightforward matter. In this part we focus on one such application, namely readability prediction.

3.1 Introduction

In the current Western society, the literacy level of the general public is often assumed to be of such a level that adults understand all texts they are confronted with on an average day. Many studies, however, have revealed that this is not the case. In the United States, for example, the 2003 National Assessment Adult Literacy showed that only 13% of adults were maximally proficient in understanding texts they encounter in their daily life. The European Commission has also been involved in extensive investigations of literacy after research had revealed that almost one in five adults in the European society lack the literacy skills to successfully function in a modern society (Wolf 2005).

Every day we are confronted with all sorts of texts, some of which are easier to process than others. Moreover, it seems that the documents which are potentially the most important for us are also the more difficult ones to process, such as mortgage files, legal texts or patient information leaflets. According to a recent OECD study, where the literacy of adults from 23 Western countries or regions was rated on a five-point scale, these specific texts genres all require a literacy level of at least four. The findings of this study for subjects from the Dutch language area show that only 12.4% of adults in Flanders and 18.2% in the Netherlands reach the two highest levels of proficiency (OECD 2013).

Readability research and the automatic prediction of readability has a very long and rich tradition (see surveys by Klare (1976), DuBay (2004), Benjamin (2012) and Collins-Thompson (2014)). Whereas superficial text characteristics leading to on-the-spot readability formulas were popular until the last decade of the previous century (Flesch 1948, Gunning 1952, Kincaid et al. 1975), recent advances in the field of computer science and natural language processing have triggered the inclusion of more intricate characteristics in present-day readability research (Si and Callan 2001, Schwarm and Ostendorf 2005, Collins-Thompson and Callan 2005, Heilman et al. 2008, Feng et al. 2010). The bulk of these studies, however, have focussed on readability as perceived by specific groups of people, such as children (Schwarm and Ostendorf 2005), second language learners (François 2009) or people with intellectual disabilities (Feng et al. 2010), and on the readability of texts from specific domains, such as the medical one (Leroy and Endicott 2011). The investigation of the readability of a wide variety of texts without targeting a specific audience, has not received much attention (Benjamin 2012).

The creation of a readability prediction system that can assess generic text material was thus a necessary end-user application which we decided to develop in the framework of this dissertation. When it comes to current state-of-the art systems, it can be observed that even though more complex features trained on various levels of complexity, have proven quite successful when implemented in a readability prediction system (Pitler and Nenkova 2008, Kate et al. 2010, Feng et al. 2010), there is still no consensus on which features are actually the best predictors of readability. As a consequence, when institutions, companies or other research disciplines wish to use readability prediction techniques, they still rely on the more outdated superficial characteristics and formulas, see for example the recent work by van Boom (2014) on the readability of mortgage terms. In addition, features that can reliably capture the highest levels of text difficulty and understanding, such as pragmatics, subtle semantics, and world knowledge are still underexplored (Collins-Thompson 2014).

This is where we hope to offer some new insights. The main focus is on exploring whether text characteristics based on deep semantic processing – in our case coreference resolution and semantic role labeling – are beneficial for automatic readability prediction.

In order to investigate this in closer detail, we needed a general evaluation corpus of Dutch generic text comprising various text genres and levels of readability, something which did not yet exist for Dutch and which we were able to derive from the SoNaR 1 corpus (cfr. Chapter 2). Since this corpus has been manually annotated with both coreference and semantic roles it is perfectly suited for our main objective in that it allows us to assess the true upper bound of incorporating deep semantic information into a readability prediction system. Next to a suitable corpus, the investigation also required a methodology to assess readability: in this respect we were the first to explore crowdsourcing as an alternative to using expensive expert labels. These first two steps are explained in Chapter 4, after which Chapter 5 describes how we built a state-of-the-art readability prediction system for Dutch and details the experimental setup that was constructed in order to investigate the added value of incorporating deep semantic knowledge. The results from the experiments are presented in Chapter 6. Parts of this research have been described and published in De Clercq et al. (2014) and De Clercq and Hoste (under review). We start by framing our work on readability prediction in the literature and focus on which text characteristics have been (successfully) implemented as features in previous work.

3.2 Related work

The central question in reading research has always been what exactly makes a particular text easy or hard to read. There seems to be a consensus that readability depends on complex language comprehension processes between a reader and a text (Davison and Kantor 1982, Feng et al. 2010). This implies that reading ease can be determined by looking at both intrinsic text properties as well as aspects of the reader. Since the first half of the 20th century, however, many readability formulas have been developed to automatically predict the readability of an unseen text based only on superficial text characteristics such as the average word or sentence length. As Bailin and Grafstein (2001) stated, these formulas appeal because they are believed to objectively and quantifiably evaluate the level of difficulty of written material without measuring characteristics of the readers.

Over the years, many objections have been raised against these traditional formulas: their lack of absolute value (Bailin and Grafstein 2001), the fact that they are solely based on superficial text characteristics (DuBay 2004, 2007, Davison and Kantor 1982, Feng et al. 2009, Kraf and Pander Maat 2009), the underlying assumption of a regression between readability and the modelled text characteristics (Heilman et al. 2008), etc. Furthermore, there seems to be a remarkably strong correspondence between the readability formulas themselves, even across different languages (van Oosten et al. 2010).

These objections have led to new quantitative approaches for readability prediction which adopt a machine learning perspective to the task. Advancements in these fields have introduced more intricate prediction methods such as Naïve Bayes classifiers (Collins-Thompson and Callan 2004), logistic regression (François 2009) and support vector machines (Schwarm and Ostendorf 2005, Feng et al. 2010, Tanaka-Ishii et al. 2010) and especially more complex features which will now be discussed in closer detail.

The **vocabulary** used in a text largely determines its readability (Alderson 1984, Pitler and Nenkova 2008). Until the beginning of the new millennium, lexical features were mainly studied by counting words, measuring lexical diversity using the type token ratio or by calculating frequency statistics based on lists (Flesch 1948, Kincaid et al. 1975, Chall and Dale 1995). In later work, a generalization over this list lookup was made by training unigram language models on grade levels (Si and Callan 2001, Collins-Thompson and Callan 2005, Heilman et al. 2007). Subsequent work by Schwarm and Ostendorf (2005) compared higher-ordered n-gram models trained on part-of-speech sequences with those using information gain and found that the latter gave the best results. To this purpose they used two paired corpora (one complex and one simplified version) to train their language models. Using the same corpora, these findings were corroborated by Feng et al. (2010) when they investigated readability targeted to people with intellectual disabilities. The above-mentioned results were thus achieved when training and testing different language models that are built on various levels of complexity. Pitler and Nenkova (2008) were the first to train unigram language models using background material complying with the genre the readability of which they were trying to assess (newspaper text). Kate et al. (2010) conducted similar experiments, but they used higher-ordered language models and normalized over document length. In subsequent work as well, this has proven a successful technique for readability prediction (Feng et al. 2010, François 2011).

In addition, the structure or **syntax**, of a text is seen as an important contributor to its overall readability. Since longer sentences have proven to be more difficult to process than short ones (Graesser et al. 2004), this traditional feature also persists in more recent work (Nenkova et al. 2010, Feng et al. 2010, François 2011). Schwarm and Ostendorf (2005) were the first to introduce more complex syntactic features based on parse trees, i.e. the parse tree height, phrase length (NP, PP, VP) and the number of subordinating conjunctions. Nenkova et al. (2010) were the first to study structural features as a stand-alone class and introduced some additional syntactic features that should be able to reflect sentence fluency. According to their findings especially features encoding the length of both sentences and phrases emerge as important readability predictors. PoS-based features, which are less hard to compute, have also been employed in previous research and have proven to be effective, too (Heilman et al. 2007), especially features based on noun and preposition word class information (Feng et al. 2010) or features representing the amount of function words present in a text (Leroy et al. 2008). Overall, Schwarm and Ostendorf's parse tree features (2005) have been reproduced a lot and were found effective when combined with n-gram modeling (Heilman et al. 2007, Petersen and Ostendorf 2009, Nenkova et al. 2010) and discourse features (Barzilay and Lapata 2008).

This brings us to a final set of features, namely those relating to **semantics**, which has been a popular focus in modern readability research (Pitler and Nenkova 2008, Feng et al. 2010, François 2011). Whereas the added value of the lexical and syntactic features has been corroborated repeatedly in the computational approaches to readibility prediction that have surfaced in the last decade, it has proven much more difficult to find corroboration for the added value of semantic features.

Capturing semantics can be done from two different angles. The first angle relates to features that are used to describe semantic concepts. The complexity and density with which concepts are included in a text can be studied by looking at the actual words that are used to describe these. Complexity was investigated in the framework of the Coh-Metrix by calculating the level of concreteness or lexical ambiguity of words against a database (Graesser et al. 2004). The validity of this approach for readability research, however, was not further investigated. Density was calculated by Feng et al. (2010) by performing entity recognition and has proven a useful feature in her work.

A second angle is to investigate how these concepts are structured within a text, for example finding semantic representations of a text or elements of textual coherence. In this respect, reference can be made to both local and global coherence, which translates to looking at the coherence between adjacent sentences (local) and then extrapolating this knowledge to reveal something about the overall textual coherence (global). This type of semantic representation can also be referred to as discourse analysis. An intuitive and straightforward way to implement this is to simply count the number of connectives included in a text based on lists or to calculate the causal cohesion by focussing on connectives and causal verbs (Graesser et al. 2004). A similar approach is to compute the actual word overlap. This word overlap was introduced without further investigations in the Coh-Metrix in three ways: noun overlap, argument overlap and stem overlap (Graesser et al. 2004). Subsequent readability research by Crossley et al. (2008) looked only at content overlap and showed this to be a significant feature. However, similar work by Pitler and Nenkova (2008) did not lead to the same conclusion. The first study to actually investigate the validity of the Coh-Metrix as a readability metric concluded that noun overlap can be indicative of causal and nominal coreference cohesion, which in turn allow to distinguish between coherent and incoherent text (McNamara et al. 2010).

More intricate methods are also available, based on various techniques. A first technique is to use latent semantic analysis (LSA). This technique was first introduced in readability research by Graesser et al. (2004) under the form of local and global LSA in the Coh-Metrix but not further investigated. The first to measure the impact of modeling local LSA for readability prediction were Pitler and Nenkova (2008); they found that the average cosine similarity between adjacent sentences was not a significant variable. Also, the validity of LSA as implemented in the Coh-Metrix could not be corroborated in the previously mentioned study by McNamara et al. (2010). François (2011) was the first to study LSA in greater detail, which seemed very helpful for his readability research for second language learners, but in more recent work his approach was criticized because of the specificity of the corpus used (Todirascu et al. 2013).

An alternative to LSA was introduced by Barzilay and Lapata (2005). They define three linguistic dimensions that are essential for accurate prediction: entity extraction, grammatical function and salience. These three dimensions are combined in the entity-grid model they propose in which all entities can be defined in a text on a sentence-to-sentence basis and where the transitions are checked for each sentence. Their main claim is that salient entities prefer prominent over non-prominent syntactic positions within a clause and are more likely to be introduced in a main clause than in a subordinate clause. Though originally devised for other research purposes, they found that the proportion of transitions in this entity grid model adds up to predicting the readability of a text in combination with the syntactic features as introduced by Schwarm and Ostendorf (2005).

Subsequent work by Pitler and Nenkova (2008) compared this entity grid model with the added value of discourse relations as annotated in the Penn Treebank (Prasad et al. 2008). They treat each text as a bag of relations rather than a bag of words and compute the log likelihood of a text based on its discourse relations and text length compared to the overall treebank. They found that these discourse relations are indeed good in distinguishing texts, especially when combined with the entity grid model. Since these discourse relations were only based on gold standard information while, in the end, a readability prediction system should be able to function automatically, Feng et al. (2010) proposed an alternative that should be able to compute this type of information. Besides entity-density and entity-grid features (cfr. supra), they introduced features based on lexical chains which try to find relations between entities (such as synonym, hypernym, hyponym, coordinate terms (siblings), etc. (Galley and Mckeown 2003)). Moreover, they incorporated coreferential inference features in order to study the actual coherence between entities. However, this study did not come to a positive conclusion for incorporating these types of features. In a follow-up study, Feng (2010) found that enlarging the corpus, which exclusively consisted of texts for primary school children before, with more diverse text material allowed for an overall better performance. However, the added value of the discourse relations to the system was still not significant.

We can conclude that the introduction of more complex linguistic features has indeed proven useful. The deep syntactic features as introduced by Schwarm and Ostendorf (2005), for example, have proven their added value in many subsequent studies. Nevertheless, there is still discussion on which of these complex features are actually the best predictors and whether it is useful to include features capturing semantics or discourse processing. While Pitler and Nenkova (2008) have clearly demonstrated the usefulness of discourse relations, the predictive power of these was not corroborated by for example Feng et al. (2010). Especially for those features requiring deep linguistic processing, a lot still has to be explored (Collins-Thompson 2014). This is exactly what will be investigated in the next chapters, which focus on readability prediction using features derived from deep semantic processing: coreference and semantic roles. We wish to explore whether these two information layers can help to make better predictions.

As a readability prediction system did not yet exist for Dutch, some first decisions had to be made regarding data selection and how this data would be assessed. In previous work we found it is difficult to make comparisons across different studies since they all use their own definition of readability and their own corpora to measure readability. Furthermore, we see that most studies focus on human judgments by, for example, people with specific disabilities or that they work with corpora of texts targeting a specific audience (mostly language learners). The work of Feng et al. (2010), for example, is very valuable due to its focus on discourse features whilst including features from previous work, but their main focus is on texts aimed at primary school students. A similar observation can be made about the work of François (2011), who investigated a wide variety of current state-of-the-art readability features, but focused on second language learners. We envisaged from the beginning to build a corpus that consists of texts adult language users are all confronted with on a daily basis. For the actual assessment, we were inspired by Dubay's (2004) vision on readability, viz. 'what is it that makes a particular text easier or more difficult to read than any other text', which means that we want to assess readability by comparing texts with each other. This definition actually summarizes the methodology we used to both assess (Chapter 4) and predict (Chapter 5) the readability of Dutch text material.

CHAPTER 4

Data collection and assessment

For the construction of a readability prediction system using supervised machine learning, three steps can be roughly distinguished. First of all, a readability corpus containing text material of which the readability will be assessed must be composed. Second, a methodology to acquire readability assessments has to be defined. Finally, based on the readability corpus and the acquired assessments, prediction tasks can be performed.

Traditionally, a readability corpus consists of reading material for language learners and is assessed by experts indicating grade levels or absolute scores. There exist almost no general domain corpora, especially not for Dutch, and other methodologies to assess readability are scarce. In this chapter, we extensively discuss the corpus we collected (section 4.1) and how this corpus has been assessed (section 4.2). For more details on the actual readability prediction system, we refer to Chapter 5.

4.1 Readability corpus

Because the main focus of readability research, until recently, has been on finding appropriate reading material for language learners, most of the existing datasets are built on underlying corpora of educational material, such as school textbooks and comparable corpora that have been collected and studied representing various reader levels (Petersen and Ostendorf 2009; Feng et al. 2010). Notable exceptions are corpora that were explicitly designed to represent an actual difference in readability based on its envisaged end-users (i.e. people with intellectual disabilities (Feng 2010)) or text genre (i.e. medical domain (Leroy and Endicott 2011)). A more general corpus, which is not tailored to a specific audience, genre or domain was assembled by the Linguistic Data Consortium (LDC) in the framework of the DARPA Machine Reading Program (Kate et al. 2010). These data, however, have not been made publicly available. At the time we started our corpus collection, for Dutch, the only large-scale experimental readability research (Staphorsius and Krom 1985, Staphorsius 1994) was limited to texts for elementary school children.

In order to build an 'unbiased' readability system, which is not targeted towards a specific audience or trained on highly specific text material only, we needed to select texts that adult language users are all confronted with on a regular, daily basis. To this purpose, we extracted 105 texts from the SoNaR 1 corpus[1]. These texts have been syntactically annotated during the LassyKlein corpus project (van Noord et al. 2013) and within the SoNaR project an additional four semantic annotation layers were added (named entities, coreferential relations, semantic roles and spatio-temporal relations).

For the envisaged assessments, we extracted snippets of between 100 and 200 words from the selected texts.[2] Most of these snippets were extracted from a larger context but are meaningful by themselves. This was objectively measured by letting two trained linguists rank 12 full texts and their 12 snippets as more difficult, less difficult or equally difficult independent of each other, with an interval of one week (i.e. 11 full text pairs and 11 snippet pairs). We opted for an interval of one week, based on the assumption that the annotators could still remember (part of) the contents of both full texts and the snippets after one week, but would have forgotten about how they ranked both groups of texts, which was also confirmed by both annotators. As the agreement between the ranking of full texts and snippets was 90.9% for the first and 100% for the second

[1] This is the same corpus that was used to perform the cross-genre robustness experiments in Chapter 2.

[2] This was necessary because we developed an interface for the actual assessments, cfr. Section 4.2, and we did not want to distract the assessors by having to scroll through large texts

annotator, we considered the snippets as viable alternatives. Please note that when, in the following sections, we refer to the texts that have been assessed, we actually mean the snippets.

The resulting readability corpus consists of texts coming from different genres in order to represent a variety of text material and presumably also of various readability levels. The *administrative* genre comprises reports and surveys or policy documents written within companies or institutions. The texts belonging to the *informative* genre can be described as current affairs articles in newspapers or magazines and encyclopedic information such as Wikipedia entries. The *instructive* genre consists of user manuals and guidelines. Finally, the *miscellaneous* genre covers other text types, such as very technical texts and children's literature. In Table 4.1, an overview is given of the number of texts and tokens included in the readability corpus together with their average sentence length.

Genre	#Documents	#Tokens	Avgsentencelen
Administrative	21	3,463	19.95
Informative	65	8,950	17.91
Instructive	8	1,108	15.98
Miscellaneous	11	1,559	19.10
Total	105	15,080	18.29

Table 4.1: Data statistics of the readability corpus.

We acknowledge that including multiple genres might influence our final training system in that it only learns to distinguish between various genres instead of various readability levels. To account for this as much as possible, we carefully tried to select texts of varying difficulty for each text genre.

$$Douma = 207 - 0.93 \cdot avgsentencelen - 77 \cdot avgnumsyl \tag{4.1}$$

$$Brouwer = 195 - 2 \cdot avgsentencelen - 67 \cdot avgnumsyl \tag{4.2}$$

$$Flesch = 206.835 - 1.015 \cdot avgsentencelen - 84.6 \cdot avgnumsyl \tag{4.3}$$

Although we are fully aware of the shortcomings of the existing readability formulas, we confronted our intuitive selection with the output of two classical readability formulas designed for Dutch so as to objectify our manual selection as much as possible.

The formulas we used, are the Flesch-Douma formula given in (4.1) (Douma 1960) and the Leesindex Brouwer given in (4.2) (Brouwer 1963). Both are adaptations of the well-known English Flesch reading ease formula given in

(4.3) (Flesch 1948). As can be seen, these formulas are based on shallow text characteristics such as the average sentence length (*avgsentencelen*) and the average number of syllables per word (*avgnumsyl*) in a particular text. The latter was calculated for Dutch by using a classification-based syllabifier (van Oosten et al. 2010).

Fig. 4.1 presents an overview of the readability scores each formula assigned to the texts in each of the genres. It should be noted that both formulas scale from 0 to circa 120 and that their results are inversely proportional, i.e. the lower the score, the more difficult the text is. Looking at the boxplots of both formulas, we immediately notice that the median values across the different genres are all quite close to each other (in between 25-40 for Brouwer and 35-50 for Douma). This indicates that, overall, it seems as though both formulas consider the selected texts to be positioned near the more difficult end of the readability continuum. This could be expected since we envisaged to select texts for an adult audience, thus assuming some level of education. However, because of the large spread, both in terms of difference between the minimum and maximum values and the interquartile range (difference between third and first quartile) we can be confident that our corpus does already represent various levels of readability per genre.

4.2 Assessments

Deciding how readability will be assessed is not a trivial task and there exists no consensus on how this should be done. One manner is to perform cloze-tests where readers are confronted with a texts from which some words are deleted and asked to fill in the blanks. Cloze, however, is a subjective evaluation that mirrors the language ability and background information of the person taking the test (Samuels et al. 1988). Today, most educators recognize that cloze procedures are more suitable to assess readers' abilities than to measure the actual readability of a text. In modern readability research, we see that most readability datasets consist of graded passages, i.e. texts have received a grade level or absolute difficulty score, typically assigned by experts (Collins-Thompson 2014).

Consulting these experts or language professionals, is a technique which is both time- and money-consuming. In NLP, many annotation tasks suffered from the same problems, which might be an explanation for the fact that in recent years, we have noticed an increasing success of cheaper and non-expert contributors over the Web, also known as crowdsourcing (Sabou et al. 2012). Thanks to web

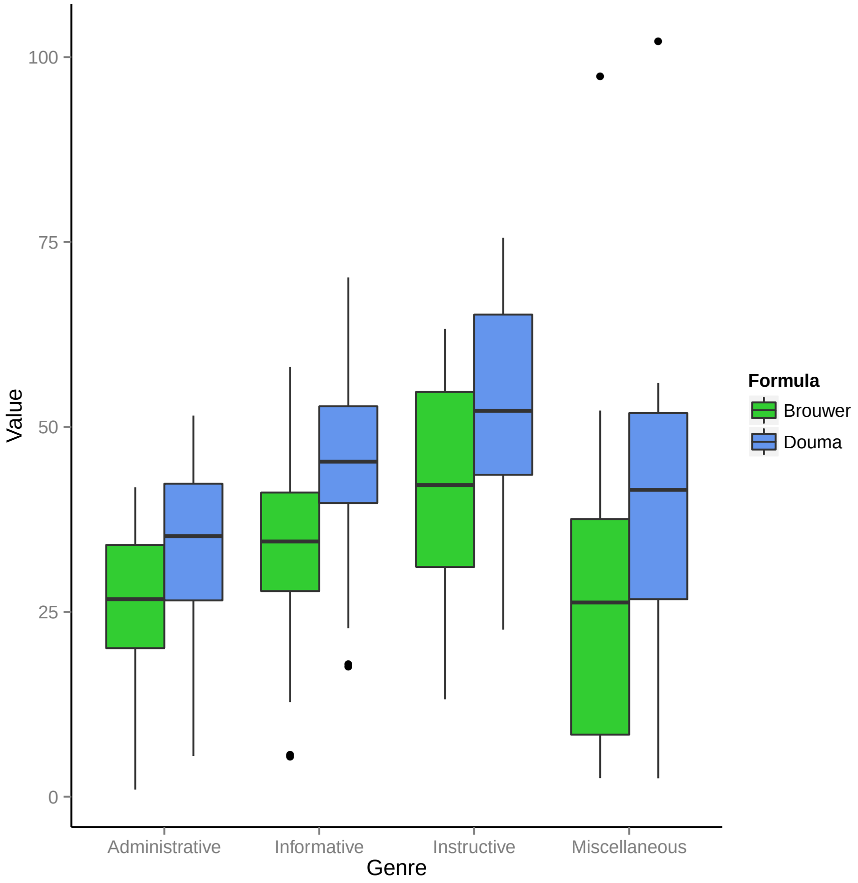


Figure 4.1: Box-and-whisker plots representing the scores of the Brouwer and Douma formula on the different genres. The maximum and minimum values in the dataset (except for the outliers) are represented as the upper end of the top whisker and the lower end of the bottom whisker. The first and third quartile are displayed as the bottom and the top of the box, the median as a horizontal stroke through the box. The outliers – indicated by a circle – are the scores which lie more than 3 times outside the interquartile range (IQR).

services such as CrowdFlower[3] or Amazon Mechanical Turk[4], this annotation technique has become very popular. An important prerequisite for using crowdsourcing is that using non-expert labels for training machine learning algorithms should be as effective as using gold-standard annotations from experts. Snow et al. (2008), for example, demonstrated the effectiveness of using Turkers for a variety of NLP tasks and found that for many tasks only a small number of non-expert annotations per item was necessary to equal the performance of an expert annotator. Bhardwaj et al. (2010), however, showed that for their more complex task of word sense disambiguation a small number of trained annotators were superior to a larger number of untrained Turkers. Nevertheless, crowdsourcing has become common practice and guidelines and best practices have been set up to streamline the process (Sabou et al. 2014).

The task of assigning readability assessments to texts is quite different from labeling tasks where a set of predefined guidelines have to be followed. Readability assessment remains largely intuitive, even in cases where annotators are instructed to pay attention to syntactic, lexical or other levels of complexity. But then again, this lack of large sets of guidelines might be another motivation to use crowdsourcing instead. This is why we decided to explore two different methodologies to collect readability assessments for our corpus, viz. a more classical expert labeling approach, in which we collect assessments of language professionals (4.2.1), and a lightweight crowdsourcing (4.2.2) approach. In our approach, all texts in the corpus are compared to each other by different people and by using different comparison mechanisms. By collecting multiple assessments per text, we aim to level out the readers' background knowledge and attitudes as much as possible. In this way, we hypothesize that a crowdsourcing approach could be a viable alternative to gathering expert labels.

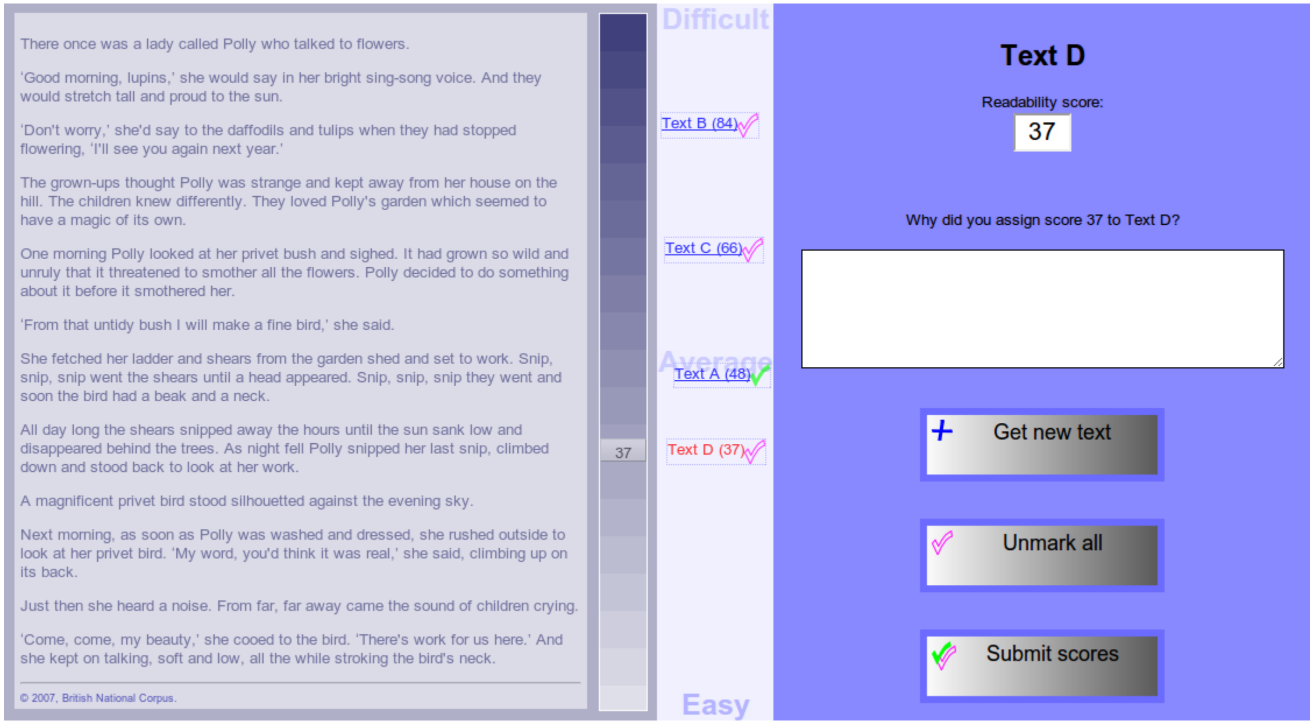
4.2.1 Expert readers

With the *Expert Readers* application[5] (see Fig. 4.2), we envisaged a more traditional approach to readability assessment in which experts rate the level of readability of a given text. The interface allows the assessors to rank all texts according to their perceived degree of readability. Through this ranking setup, the number of pairwise comparisons being performed grows quadratically with each assessed text.

[3] http://www.crowdflower.com/

[4] https://www.mturk.com/

[5] The Expert Readers application is accessible at the password-protected url http://lt3.ugent.be/tools/expert-readers-nl/.



There once was a lady called Polly who talked to flowers.

'Good morning, lupins,' she would say in her bright sing-song voice. And they would stretch tall and proud to the sun.

'Don't worry,' she'd say to the daffodils and tulips when they had stopped flowering, 'I'll see you again next year.'

The grown-ups thought Polly was strange and kept away from her house on the hill. The children knew differently. They loved Polly's garden which seemed to have a magic of its own.

One morning Polly looked at her privet bush and sighed. It had grown so wild and unruly that it threatened to smother all the flowers. Polly decided to do something about it before it smothered her.

'From that untidy bush I will make a fine bird,' she said.

She fetched her ladder and shears from the garden shed and set to work. Snip, snip, snip went the shears until a head appeared. Snip, snip, snip they went and soon the bird had a beak and a neck.

All day long the shears snipped away the hours until the sun sank low and disappeared behind the trees. As night fell Polly snipped her last snip, climbed down and stood back to look at her work.

A magnificent privet bird stood silhouetted against the evening sky.

Next morning, as soon as Polly was washed and dressed, she rushed outside to look at her privet bird. 'My word, you'd think it was real,' she said, climbing up on its back.

Just then she heard a noise. From far, far away came the sound of children crying.

'Come, come, my beauty,' she cooed to the bird. 'There's work for us here.' And she kept on talking, soft and low, all the while stroking the bird's neck.

© 2007, British National Corpus.

Text D

Readability score: 37

Why did you assign score 37 to Text D?

Figure 4.2: A screenshot of the *Expert Readers* web application.

The experts can express their opinion by ranking texts on a scale of 0 (easy) to 100 (difficult), which allows them to compare texts while at the same time assigning absolute scores.

For the assessment of our corpus, we specifically aimed at people who are professionally involved with language, thus following a more traditional data collection methodology. The experts were asked to assess the readability for language users in general. We deliberately did not ask more detailed questions about certain aspects of readability, because we wanted to avoid influencing the text properties experts pay attention to. Neither did we inform the experts in any way on how they should judge readability. Any presumption about which features should be regarded as important readability indicators was thus avoided.

However, in order to have some idea about their assessment rationale the experts were offered the possibility to motivate or to comment on their assessments via a free text field in the interface. The experts were given access to the website, could assess at their own pace and were free to decide when to submit an annotation session. Every submitted session was stored in a batch. Our pool of experts consisted of 36 teachers, writers and linguists – all native speakers of Dutch. In total, these experts contributed 2,564 text rankings.

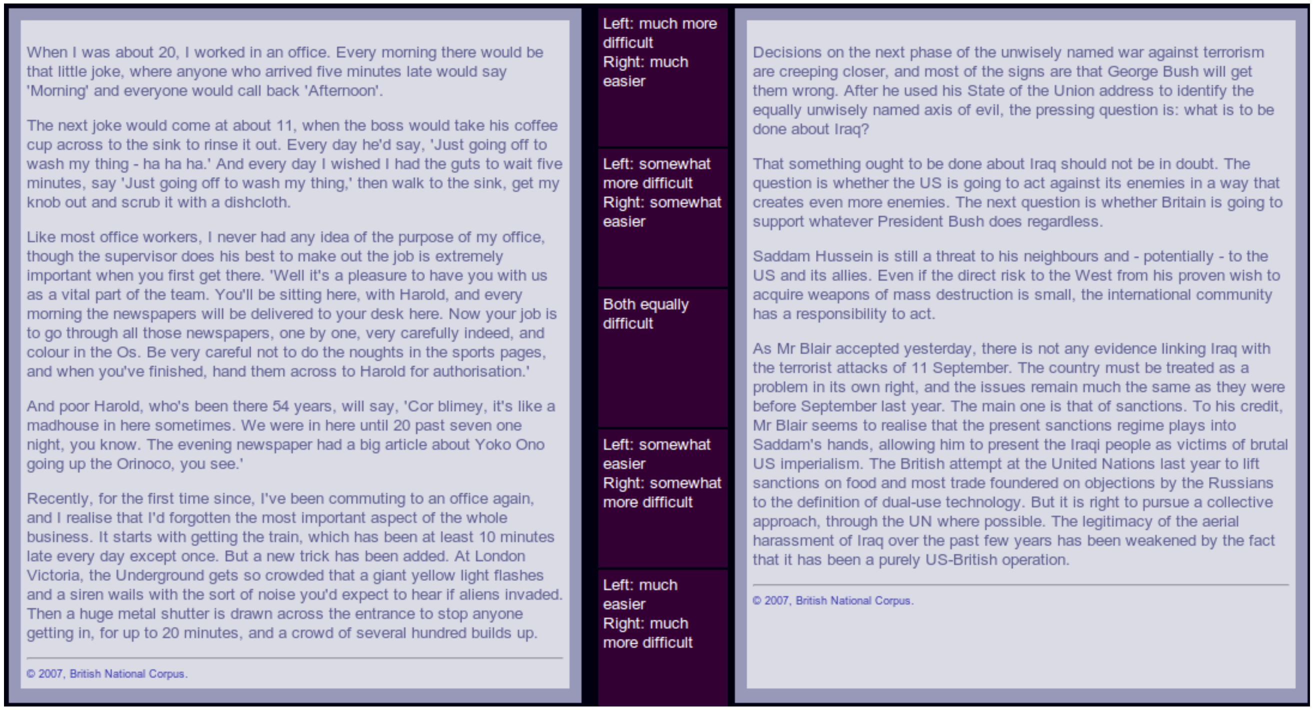
4.2.2 Crowdsourcing

Our crowdsourcing application, called *Sort by Readability*[6], was designed to be used by as many people as possible and users were not required to provide personal data or to log in. A screenshot of the crowdsourcing application is shown in Fig. 4.3. Two texts are displayed simultaneously and the user is asked to tick one of the statements in the middle column, corresponding to the five-point scale in Table 4.2.

Acronym	**Meaning**
LME	left text much easier
LSE	left text somewhat easier
ED	both texts equally difficult
RSE	right text somewhat easier
RME	right text much easier

Table 4.2: Five-point scale for the assessment of the difference in readability between two texts

[6] The Sort by Readability application can be accessed through the following url: http://lt3.ugent.be/tools/sort-by-readability-nl/.



When I was about 20, I worked in an office. Every morning there would be that little joke, where anyone who arrived five minutes late would say 'Morning' and everyone would call back 'Afternoon'.

The next joke would come at about 11, when the boss would take his coffee cup across to the sink to rinse it out. Every day he'd say, 'Just going off to wash my thing - ha ha ha.' And every day I wished I had the guts to wait five minutes, say 'Just going off to wash my thing,' then walk to the sink, get my knob out and scrub it with a dishcloth.

Like most office workers, I never had any idea of the purpose of my office, though the supervisor does his best to make out the job is extremely important when you first get there. 'Well it's a pleasure to have you with us as a vital part of the team. You'll be sitting here, with Harold, and every morning the newspapers will be delivered to your desk here. Now your job is to go through all those newspapers, one by one, very carefully indeed, and colour in the Os. Be very careful not to do the noughts in the sports pages, and when you've finished, hand them across to Harold for authorisation.'

And poor Harold, who's been there 54 years, will say, 'Cor blimey, it's like a madhouse in here sometimes. We were in here until 20 past seven one night, you know. The evening newspaper had a big article about Yoko Ono going up the Orinoco, you see.'

Recently, for the first time since, I've been commuting to an office again, and I realise that I'd forgotten the most important aspect of the whole business. It starts with getting the train, which has been at least 10 minutes late every day except once. But a new trick has been added. At London Victoria, the Underground gets so crowded that a giant yellow light flashes and a siren wails with the sort of noise you'd expect to hear if aliens invaded. Then a huge metal shutter is drawn across the entrance to stop anyone getting in, for up to 20 minutes, and a crowd of several hundred builds up.

© 2007, British National Corpus.

Left: much more difficult Right: much easier
Left: somewhat more difficult Right: somewhat easier
Both equally difficult
Left: somewhat easier Right: somewhat more difficult
Left: much easier Right: much more difficult

Decisions on the next phase of the unwisely named war against terrorism are creeping closer, and most of the signs are that George Bush will get them wrong. After he used his State of the Union address to identify the equally unwisely named axis of evil, the pressing question is: what is to be done about Iraq?

That something ought to be done about Iraq should not be in doubt. The question is whether the US is going to act against its enemies in a way that creates even more enemies. The next question is whether Britain is going to support whatever President Bush does regardless.

Saddam Hussein is still a threat to his neighbours and - potentially - to the US and its allies. Even if the direct risk to the West from his proven wish to acquire weapons of mass destruction is small, the international community has a responsibility to act.

As Mr Blair accepted yesterday, there is not any evidence linking Iraq with the terrorist attacks of 11 September. The country must be treated as a problem in its own right, and the issues remain much the same as they were before September last year. The main one is that of sanctions. To his credit, Mr Blair seems to realise that the present sanctions regime plays into Saddam's hands, allowing him to present the Iraqi people as victims of brutal US imperialism. The British attempt at the United Nations last year to lift sanctions on food and most trade foundered on objections by the Russians to the definition of dual-use technology. But it is right to pursue a collective approach, through the UN where possible. The legitimacy of the aerial harassment of Iraq over the past few years has been weakened by the fact that it has been a purely US-British operation.

© 2007, British National Corpus.

Figure 4.3: A screenshot of the *Sort by Readability* web application.

After clicking one of the options, the text pair and its corresponding assessment are added to the database and two new randomly selected texts appear. To avoid that respondents click on one of the buttons too soon, i.e. without reading the texts, the buttons are disabled during a few seconds after the appearance of a new text pair.

The only instructions respondents are given, are the following two sentences on the landing page of the application:

> "Using this tool, you can help us compose a readability corpus. You are shown two texts for which you can decide which is the more difficult and which is the easier one."

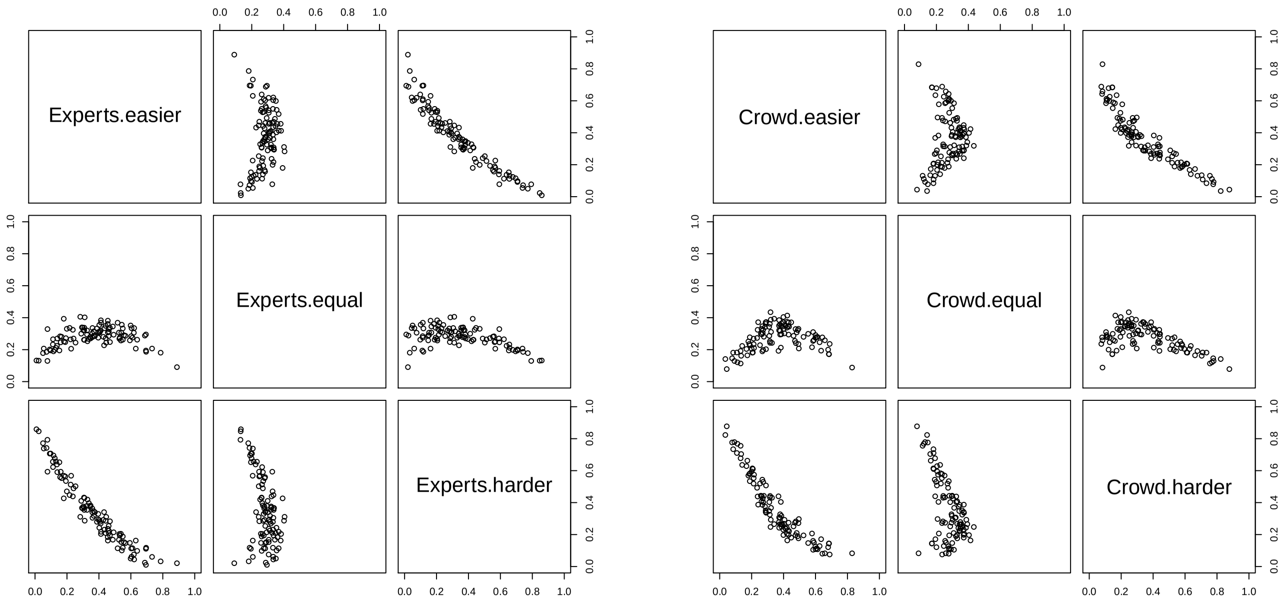
As was done for the experts, we gave no further instructions because we did not want to influence anyone on how to perceive readability. Since we deliberately chose to keep the crowdsourcing tool open to everyone, we do not know who performed the assessments. In the start-up phase, the tool was widely advertised among friends, family, students, etc., which might have caused a bias towards more educated labelers. But evidently, we do not have a clear view on the identity and background of the people who provided assessments. We can state with certainty, however, that the users of the crowdsourcing application differed from the experts selected for the first application (Section 4.2.1). In total, 11,038 comparisons were performed and the number of assessments per text pair varies from 1 to 8.

4.2.3 Experts versus crowd

In order to compare the information collected with both applications and perform readability prediction experiments (Chapter 5), all data had to be converted to assessed text pairs. The experts batches were transformed into a dataset comprising 23,908 text pairs, and the crowd dataset consisted of 11,038 text pairs.[7]

A comparison of both datasets revealed some interesting correlations as illustrated in Figure 4.4. The proportions with which each text has been assessed as easier, equally readable or harder for both the experts and crowd data set are shown in Fig. 4.4. Since the lower left triangle and upper right triangle in both figures present the same information, we will limit the discussion to the lower left triangle. Each dot in the figures represents one text, so every plot in both figures represents the 105 assessed texts.

[7] For more details on how the experts data was transformed we refer to De Clercq et al. (2014).



(a)

(b)

Figure 4.4: Proportion of times each text was assessed as easier, equally difficult or harder than any other text: (a) for the *Sort by Readability* data and (b) for the *Expert Readers* data. In both figures, the plots in the lower left triangle are transposed versions of those in the upper right triangle (the x and y axes are switched). They therefore present the same information.

If we take for example text 6, we see that this text has been assessed in our experts dataset 0.88 times as easier, 0.09 times as equally difficult and 0.02 times as more difficult than any other text. In our crowd dataset the same text has been assessed 0.83 times as easier, 0.09 times as equally difficult and 0.08 times as more difficult than any other text. These scores are visualized by the rightmost dot in the lower left plot of subfigures a and b. Overall, we observe that all plots show great similarity for both datasets.

After calculating the Pearson correlations we found that the correlation between crowd and experts regarding the easier texts is 86%, and 90% when we look at the number of times a text was considered harder. This allows us to conclude that two very similar datasets can be derived from the applications, which means that – using our methodology – experts rank the texts from the corpus in a very similar order as the crowd does by comparing them.

4.3 Towards readability prediction datasets

The strong correlations between our experts and crowd datasets made us confident that we could combine both datasets for the actual readability prediction experiments. This led to a dataset comprising 34,946 assessed Dutch text pairs. Because this dataset still contained different assessments per text pair, it first had to be normalized[8], which was done by averaging (Equation 4.4):

$$\mathit{average} = (a_1 + a_2 + a_3, +... + a_n)/n \tag{4.4}$$

where, a_1,a_2,a_3,......,a_n is a set of assessments for one text pair. In order to be able to calculate an average value, every assessment label was assigned a corresponding value[9]. This results in the dataset as presented in Table 4.3.

There is an even distribution because every text pair has been included in both directions. This is the dataset that will be used for the readability prediction experiments in the following chapters. For all our experiments, we considered the readability prediction task as a classification task of text pairs.

A final matter that needed to be resolved before we could proceed to the actual experiments was whether our assessment techniques have actually allowed us to create a dataset truly representing various levels of readability. In order to do so,

[8] Every text should only once be compared to a different text, i.e. 105 * 104 = 10,920.

[9] The assessment label LME, for example, means that the left text is much easier that the right text which corresponds to this pair receiving the value 100 (left text minus right text, i.e. 100 - 0)

Acronym	**Meaning**	Value	Number of pairs
LME	left text much easier	100	260
LSE	left text somewhat easier	50	2782
ED	both texts equally difficult	0	4836
RSE	right text somewhat easier	-50	2782
RME	right text much easier	-100	260

Table 4.3: Total amount of text pairs for each of the five scales .

we compared all text assessments derived from our combined dataset (cfr. Table 4.3) to the scores we received for each of these texts using the traditional Brouwer and Douma formulas (see Section 4.1).

Since this required us to compare on the text level, we derived assessment statistics from our dataset comprising 10,920 text pairs. To be more precise, we indicated how many times each text was assessed as more difficult than any other text in the corpus and normalized this score to a scale from 0 to 100. In a next step, the scores that were assigned to each text by our two classical readability formulas were also normalized to the same scale. This enabled us to directly compare our assessed corpus with the Brouwer and Douma scores, as illustrated in Figure 4.5.

Looking at the jitter plots, some interesting observations can be made. First, we notice that based on our manual assessments we have obtained a dataset which is better distributed. Furthermore, we can also observe that the manual assessors consider most texts as easier than the levels assigned by the various readability prediction formulas. As we have seen, the classical formulas rate the majority of the texts as difficult to very difficult. This is because these formulas were developed to select and rate reading material for children. Clearly, they are not very suitable for scoring text material adults are confronted with on a daily basis.

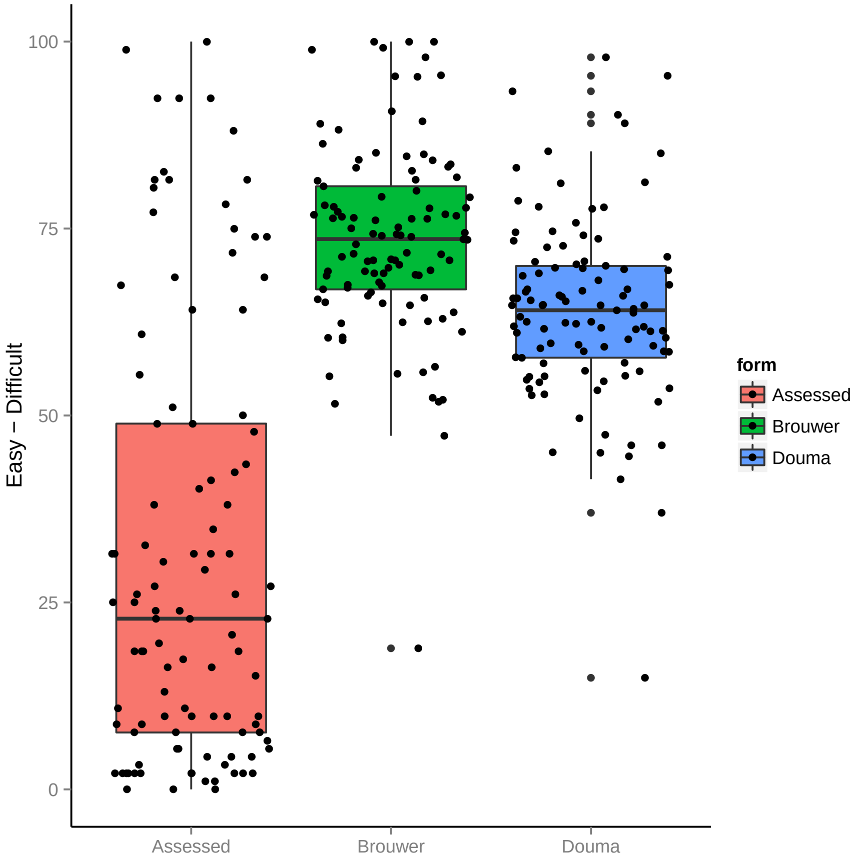


Figure 4.5: Box and jitter plots comparing the manual assessments to the classical Brouwer and Douma formulas.

CHAPTER 5

Experiments

In this chapter we present our readability prediction system and the experimental setup which should allow us to assess the added value of incorporating deep semantic information in the form of coreference and semantic role features. Basically, we perceive the task of readability prediction as a classification task.

We start by presenting the features that were derived from the texts. In Section 5.1 we describe the traditional features (used in previous readability research) as well as the novel features we added on top of this (i.e. the features derived from the coreference and semantic role labeling systems described in Chapter 2). We are mainly interested in the contribution of the two semantic layers and investigate this by first manually including or excluding these features from our classification experiments, after which we explore a wrapper-based approach to feature selection using genetic algorithms (Section 5.2), something which had not been investigated in readability research before.

5.1 Information sources

We selected the features to be implemented in our readability prediction system on the basis of both the existing literature on the topic (see the overview in Section 3.2) and the comments left by the experts participating in the readability assessment (see Section 4.2.1[1]). The scrutiny of these comments allowed us to discover some interesting tendencies with respect to which text characteristics guided their assessments most. Although the experts did not receive any guidelines on which characteristics to take into consideration when assessing readability, most assessors commented on their assessments in a similar manner. These comments can be categorized into four groups as illustrated in Figure 5.1.

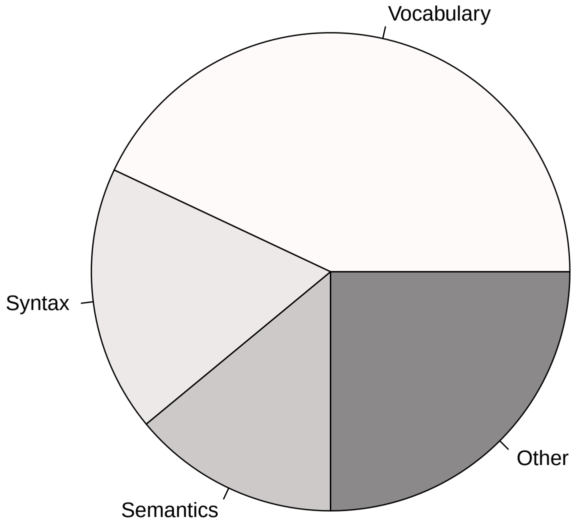


Figure 5.1: Pie chart visualizing the characteristics that were considered most important according to the 36 expert assessors that participated in our readability study.

The first class includes all comments relating to **Vocabulary** in some way or another, including comments relating to lexical familiarity (example 8) or the level of concreteness (example 9). A second class, **Syntax**, includes comments relating to syntactic constructs ranging from superficial characteristics (example 10), to complaints about more complex structures (example 11). The third

[1] All 36 experts that participated in our study did motivate their assessments via the free text field in the interface (the white box in Figure 4.2).

class groups all comments that relate to the **Semantics/Coherence** of the overall discourse and again ranges from simple (example 12) to more complex issues (example 13). Finally, the **Other** class contains all those comments that could not be grouped under a certain linguistic category (example 14).

(8) NL: Voor specialisten zal dit een gemakkelijk te lezen tekst zijn, maar een leek zal niet altijd goed kunnen volgen als hij de betekenis van de specifieke woorden niet kent.
EN: For specialists this text will be easy to read, but a layman will not always be able to follow if he/she does not know the meaning of specific words.

(9) NL: vrij hoog abstractieniveau, in thema en woordenschat ("overheidsinstrumenten", "financieringsproducten","investeringsmiddelen").
EN: high level of abstraction, in both theme and vocabulary ("government instruments", "products to finance", "investments products").

(10) NL: tekst vergt veel concentratie, te lange zinnen
EN: text requires much concentration, too long sentences

(11) NL: Pfffff, beginnen met een bijzin, zinnen tussen haakjes, passiefvormen, constructies als "in dat geval is het verstandig de hulp in te roepen van"... De ene bijzin wordt bovenop de andere geplaatst.
EN: Pffff, starting with a subordinated clause, sentences in between brackets, passive constructions, constructions such as "in that case it could be wise to consult"... One subordination on top of the other.

(12) NL: er staan bijna geen verbindingswoorden, waardoor de link niet altijd gemakkelijk gelegd kan worden tussen twee zinnen.
EN: almost no connectives are used, as a result of which it is not always easy to see the link between two sentences.

(13) NL: ... verwijzingen zijn onduidelijk: 3e paragraaf verwijst naar wat?, die komma-zin is verwarrend. 4e paragraaf: voorstel=hoe zo, welk voorstel? wat valt er toe te wijzen, een voorstel? Laatste zin: "die middelen": welke?
EN: ... references are unclear: to what does the third paragraph refer?, that comma-sentence is confusing. fourth paragraph: proposition = come on, what proposition? what is there to attribute, a proposition? Last sentence: "those means": which ones?

(14) NL: Toch moet ik de tekst twee keer lezen vooraleer ik echt weet waarover het gaat.
EN: Still, I had to read the text twice before I really knew what it was about.

According to our expert assessors, the vocabulary of a text seems to be the most important obstructor or facilitator of text readability. It accounts for almost half of all comments (43%). This might indicate that lexical features are indeed crucial when trying to predict readability. However, the syntactic and semantic aspects of a text, good for 18% and 14% of the comments, respectively, should not be ignored either. Referring back to our literature overview (Section 3.2) we observed the popularity of language models or other more complex lexical features for readability prediction using machine learning (Pitler and Nenkova 2008, Kate et al. 2010). Besides complex lexical features, the deep syntactic features as introduced by Schwarm and Ostendorf (2005) have also proven their added value, whereas the importance of semantic and discourse processing has proven more difficult to corroborate (Pitler and Nenkova 2008, Feng et al. 2010).

We will now give an overview of the novel features that were introduced in our readability prediction system and with which we explore the added value of adding deep semantics to readability prediction. We start by explaining the features we derived from our two semantic layers under consideration – coreference and semantic roles – after which other features, commonly used in previous readability research, are introduced. This is the first time such an elaborate feature vector was encoded for Dutch readability prediction.

5.1.1 Coreference features

During reading, a reader has to build a coherent semantic representation of a text that allows him or her to actively retrieve and assess the previously processed information (Feng 2010). Reading and understanding a text requires a person to identify each linguistic unit, understand what this unit refers to and link those units that are coreferent. Coreference takes place between discourse entities which may or may not correspond to something in the real world (Webber 1978). Indicating coreferential relations in a text should thus allow us to indicate how complex (i.e. how many different entities are referred to), but also how structured and coherent (i.e. how easy is it for the reader to assess previously stored information) a particular text is.

We have gold-standard coreference annotations available for all 105 texts in our readability corpus and we wish to compare features derived from this gold information to those extracted from automatically annotated texts using the COREA tool. To this purpose, the COREA tool was retrained on a large collection of annotated text material, i.e. 500,000 words from the SoNaR 1 corpus which comprises six different text genres. In Chapter 2, we explained that training on such a comprehensive and varied dataset is the best approach when you want to automatically label a variety of text material. We did make sure, however, that our training dataset did not contain any of the 105 texts occurring in our readability corpus.

We extracted all noun phrases (common, proper nouns and pronouns) referring to the same person or object and built coreference chains. We are aware that limiting the text length to snippets of between 100 and 200 words might have influenced the number of coreferential relations present within our readability corpus. However, based on the gold-standard annotations, we found that our corpus includes texts which contain from one to eleven coreferential chains, which hints at a large variability. Five features, as listed in Table 5.1 were derived from this information.

Group	Feature
coref	number of coreference chains per document (*numchains*)
	average chain span per document (*chainspan*)
	number of large chain spans within a document (*largespan*)
	average number of coreferential NPs within a document (*corefs*)
	average unique coreferential NPs within a document (*unicorefs*)

Table 5.1: Table listing the five coreference features together with their abbreviation.

An example of a coreferential chain extracted from a text from our corpus is presented in Figure 5.2. The two numbers on the left represent the chain number and the token index indicating the end of a particular noun phrase. The two phrases on the right represent the head of the noun phrase and the actual noun phrase. All NPs are sorted in ascending order and the first phrase is the antecedent to which all other phrases refer back. We observe that ten references are made to the first NP ‘Astrid Sofia Lovisa Thyra’ which implies that the text helps the reader to build a coherent semantic representation.

Based on this chain and inspired by the work of Feng (2010), we implemented three chain features: the number of coreferential chains present in a text, the average chain length, and we also counted how many coreferential chains span more than half of the text. Applied to our example, the total number of chains in this text amounts to two which means that not many extra-linguistic entities are introduced and discussed in this text. The span of our example chain is 116 words (120 - 4) and since the text length of this particular document is 126

2	4	astrid_sofia_lovisa_thyra	Astrid Sofia Lovisa Thyra
2	19	prinses	prinses van België
2	23	hertogin	hertogin van Brabant
2	27	prinses	prinses van Zweden
2	41	dochter	de dochter van prins Karel van Zweden en prinses Ingeborg van Denemarken
2	45	haar	haar
2	57	zij	zij
2	69	koningin	koningin der Belgen
2	72	zij	Zij
2	104	zij	Zij
2	120	moeder	de moeder van de latere koningen Boudewijn en Albert II en groothertogin Josephine-Charlotte van Luxemburg

Figure 5.2: Figure representing a coreferential chain.

words, the example chain is spanning more than half of the document and can be considered a large chain span. The other chain in this text spanned only nine words, which leaves us with an average chain span of 62.5 words.

In order to have some idea of the number of inferences that are required within a text, we also counted the number of coreferential NPs within a document and the number of unique coreferential NPs. When a text contains more different NPs referring back to the same antecedent, this requires more mental actions from the reader in that he or she has to infer more. On the other hand, in some cases, it might be more readable to use an additional lexical NP instead of another pronoun, especially when a certain pronoun has many possible antecedents. The total number of unique coreference NPs should grasp this aspect while also indicating some level of world knowledge.

The chain features have been tested before in the work of Feng (2010). She found that none of these features possesses a high predictive power for readability research, which she attributed to the low performance of the underlying coreference system on which the features were based. In Chapter 2, we learned that the best score received using the COREA system was a MUC F-measure of 51.4%.

5.1.2 Semantic role features

In order to compute a semantic representation for an entire sentence, semantic roles are useful. With semantic roles one can relate syntactic complements to semantic arguments, resulting in predicate-argument structures. Semantic roles indicate *Who did what to whom, when, where and how.* The added value of semantic roles has not been investigated in readability research before. As explained in Chapter 2, we use the semantic roles as indicated in PropBank (Palmer et al. 2005). We have gold-standard semantic roles annotations available for all 105 texts in our readability corpus and we wish to compare features derived from this gold information to those extracted from automatically annotated texts using the SRRL tool. To this purpose, the SSRL tool was trained on all five text genres present within the SoNaR corpus. Again, we excluded the 105 texts that also occurred in our dataset from the training base.

For other NLP tasks using the PropBank approach, semantic roles have proven interesting to disambiguate the senses of verbs and the roles of their arguments. However, we decided not to include such detailed features because we are more interested in the overall readability of a text. In order to determine how many agents or modifiers a particular text contains, we calculated the average number of arguments and modifiers and the average occurrence of every possible PropBank label (Palmer et al. 2005) per sentence as presented in Table 5.2

We believe that statistical information derived from the number of semantic roles present within a text can tell something more about its level of complexity. A text containing few semantic roles might hint at a high level of complexity because they contain many nominalisations or passive constructions. On the other hand, an abundant number of the two most prominent roles: Arg0, Arg1 might indicate that the text contains a lot of simple sentences (example 15). However, the same high level of the most prominent arguments together with other argument such as Arg2, Arg3 and Arg4 might hint at more complex structures (example 16). Some modifiers might give more context and thus render the text more readable, such as modifiers of manner, location or time (example 17). However, others could render the text more complex, such as modifiers of purpose, modality or negation (example 18).

(15) NL: Lila het vosje spitst de oren.
EN: Lila the fox pricks its ears up.
Roles: *Lila het vosje* (Arg0), *spitst* (PRED, prick.02), *de oren* (Arg1).

(16) NL: De ontwikkeling van PPS maakt in ieder geval deel uit van een bredere evolutie waarbij de rol van de overheid in de economie verandert van rechtstreekse speler naar organisator, regelgever en controleur.
EN: The evolution of PPS consists certainly of a broader evolution in which the governmental role in the economy are changing from active player to organizer, lawmaker and controller.
Roles: *De ontwikkeling van PPS* (Arg1), *maakt deel uit* (PRED, consist.01), *in ieder geval* (ArgM-ADV), *van een bredere ...* (Arg2).

Group	feature
semsrl	average number of arguments (*args*)
	average number of modifiers (*mods*)
	average number of external arguments, proto-agents (*arg0*)
	average number of internal arguments, proto-patients (*arg1*)
	average number of indirect objects, beneficiaries, etc. (*arg2*)
	average number of start points, instruments, etc. (*arg3*)
	average number of end points (*arg4*)
	average number of adverbials (*modadv*)
	average number of causal clauses (*modcau*)
	average number of directionals (*moddir*)
	average number of discourse markers (*moddis*)
	average number of extent markers (*modext*)
	average number of locatives (*modloc*)
	average number of manner markers (*modmnr*)
	average number of modals (*modmod*)
	average number of negations (*modneg*)
	average number of purpose markers (*modpnc*)
	average number of secondary predicates (*modprd*)
	average number of reciprocals (*modrec*)
	average number of temporals (*modtmp*)

Table 5.2: Table representing the 20 semantic role features included in our system.

de rol van de overheid in de economie (Arg1), *verandert* (PRED, change.01), *van rechtstreekse speler* (Arg3), *naar organisator, regelgever en controleur* (Arg2).

(17) NL: De correspondent van de The Wall Street Journal verdween op 23 januari in de Pakistaanse stad Karachi.
EN: The correspondent of The Wall Street Journal disappeared on January 23rd in Karachi, Pakistan.
Roles: *De correspondent van de The Wall Street Journal* (Arg1), *verdween* (PRED, disappear.01), *op 23 januari* (ArgM-TMP), *in de Pakistaanse stad Karachi* (ArgM-LOC).

(18) NL: Bij monde van senator Jeannine Leduc laat de VLD voor het eerst openlijk blijken dat ze het eigen voorstel voor een spijtoptantenregeling heeft bevroren opdat de PS het migrantenstemrecht niet zou goedkeuren.
EN: Through senator Jeannine Leduc, VLD for the first time openly showed to have frozen its own proposal for the pentito arrangement, lest the PS pass immigrant voting.

> Roles: *Bij monde van senator Jeannine Leduc* (ArgM-ADV), *de VLD* (Arg0), *laat blijken*, (PRED, show.01), *voor het eerst* (ArgM-ADV), *openlijk* (ArgM-MNR), *dat ze het eigen voorstel voor een spijtoptantenregeling heeft bevroren* (Arg1), *opdat de PS het migrantenstemrecht niet zou goedkeuren* (ArgM-PNC).
> *ze* (Arg0), *het eigen voorstel voor een spijtoptantenregeling* (Arg1), *bevroren* (PRED, freeze.01).
> *de PS* (Arg0), *het migrantenstemrecht* (Arg1), *niet* (ArgM-NEG), *zou* (ArgM-MOD), goedkeuren (PRED, approve.01).

Especially the interplay between syntax and semantics is interesting here. With these semantic role features we try to grasp some parts of the semantics of the structure whereas we will also include deep syntactic features. We hypothesize that exactly the combination of these two feature groups should allow us to better predict readability.

5.1.3 Other features

Besides these two important feature groups, we implemented various lexical, syntactic and other semantic features used in previous research. Furthermore, we also decided to integrate the more ‘traditional’ lexical and syntactic features – those that are used in the classical readability formulas – as a separate group because, in recent work, these have proven good predictors of readability in addition to the NLP-inspired features (Pitler and Nenkova 2008, François 2011). In total, we encoded an additional 61 features, which were all computed on the document level using state-of-the-art text processing tools. These are now discussed in detail whereas a schematic overview can be found in Table 5.4.

Traditional features

We included four length-related features that have proven successful in previous work (Nenkova et al. 2010, Feng et al. 2010, François and Miltsakaki 2012): the average word and sentence length, the ratio of long words in a text (i.e. words containing more than 3 syllables), and the percentage of polysyllable words. We also incorporated two traditional lexical features. On the one hand the percentage of words also found in a Dutch word list with a cumulative frequency of 77%[2]. On the other hand we also calculated the type token ratio – TTR –

[2] The list is based on a list ordered by descending frequency in a large newspaper corpus, i.e. the ‘27 Miljoen Woorden Krantencorpus 1995’ which is available through the HLT agency at http://tst.inl.nl/en/producten.

(Equation 5.1) to measure the level of lexical complexity within a text. All these features were obtained after processing the text with a state-of-the-art Dutch preprocessor (Frog (van den Bosch et al. 2007)) and a designated classification-based syllabifier (van Oosten et al. 2010).

$$\text{type token ratio} = \frac{\text{types}}{\text{total number of words}} \tag{5.1}$$

Lexical features

Since we tried not to have presuppositions about the various levels of complexity in our corpus, we decided to build a generic language model for Dutch based on a subset of the SoNaR corpus (Oostdijk et al. 2013). This subset contains only newspaper, magazine and wikipedia material and should qualify as a generic representation of standard written Dutch. The language model was built up to an order of 5 ($n = 5$) with Kneser-Ney smoothing using the SRILM toolkit (Stolcke 2002). As features we calculated the perplexity of a given text when compared to this reference data and also normalized this score by including the document length, as seen in Kate et al. (2010).

Besides these n-gram models, which have proven strong predictors of readability in previous work (Kate et al. 2010, Feng et al. 2010, François 2011), we also introduced two other lexical features which were calculated using the same reference corpus. Inspired by terminological work, we included the Term Frequency-Inverse Document Frequency, aka TF-IDF and the Log-Likelihood ratio of all terms included in a particular text.

TF-IDF originates from information retrieval and measures the relative importance or weight of a word in a document (Salton 1989). We calculated TF-IDF for all terms in the readability corpus and to calculate the idf we enlarged the readability corpus with all texts of our reference corpus. Given a document collection D, a word w, and an individual document d in D,

$$W_d = f_{w,d} \cdot \log(|D|/f_{w,D}) \tag{5.2}$$

where $f_{w,d}$ equals the number of times w appears in d, $|D|$ is the size of the corpus and $f_{w,D}$ equals the number of documents in D in which w appears (Berger et al. 2000). Calculating TF-IDF should thus enable us to extract those specific words in our texts that have much lower frequencies in the balanced background corpus. In short, the mean TF-IDF value of all the words in a large corpus estimates the mean importance of a word in any text.

	First Corpus	Second Corpus	Total
Frequency of word	a	b	a+b
Frequency of other words	c-a	d-b	c+d-a-b
Total	c	c	c+d

Table 5.3: Contingency table to calculate Log-Likelihood.

The Log-Likelihood ratio discovers keywords which differentiate between corpora, in our case the reference and the input text/corpus. We first produced a frequency list for each corpus and calculated the Log-Likelihood statistic for each word in the two lists. This was done by constructing a contingency table (see Table 5.3), where c represents the number of words in the reference corpus and d corresponds to the number of words in our corpus. The values a and b are known as the observed values (O).

Next, the expected value for each word is calculated as follows:

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i} \tag{5.3}$$

where N corresponds to the total number of words in the corpus and i to the single words. The observed values correspond to the real frequency of a single word i in the corpus. So, for each $word_i$, the observed value O_i is used to calculate the expected value. Applying this formula to our contingency table (with N_1=c and N_2=d) results in:

$$E_1 = c \cdot (a+b)/(c+d) \tag{5.4}$$

$$E_2 = d \cdot (a+b)/(c+d) \tag{5.5}$$

Finally, the resulting expected values are used for the calculation of the Log-Likelihood (LL):

$$-2ln\lambda = 2\sum_i O_i ln \left(\frac{O_i}{E_i}\right) \tag{5.6}$$

which equates to:

$$LL = 2 \cdot \left(a \cdot \log\left(\frac{a}{E_1}\right)\right) + \left(b \cdot \log\left(\frac{b}{E_2}\right)\right) \tag{5.7}$$

More information about the calculation of the expected values and Log-Likelihood can be found in Rayson and Garside (2000). Since the reference corpus models usual everyday Dutch language, the intuition here is that texts with an overall unnatural use of words will be detected by the Log-Likelihood ratio.

Syntactic features

We incorporated two types of syntactic features: a shallow level where all feature are computed based on PoS-tags and a deeper level based on dependency parsing. We included 25 shallow features, inspired by Feng et al. (2010), relating to the five main part-of-speech classes: nouns, adjectives, verbs, adverbs and prepositions. For each class, we indicated their average frequency on the text (e.g. total number of nouns normalized over text length) and sentence level (e.g. number of nouns averaged over sentence length). In addition, we also calculated the average PoS types (e.g. number of unique nouns) on both the document and sentence level and the ratio of unique PoS types over the total number of unique words within a document. To finish, we calculated two additional features, the average number of content and function words within a text (Leroy et al. 2008). For these calculations, the same preprocessor tool was used as mentioned before (i.e. Frog).

For the deep syntactic features, we incorporated the parse tree features as first introduced by Schwarm and Ostendorf (2005) and that have proven successful in many other studies (Pitler and Nenkova 2008, Petersen and Ostendorf 2009, Nenkova et al. 2010, Feng et al. 2010). We calculated the parse tree height, the number of subordinating conjunctions and the ratio of the noun, verb and prepositional phrases. As an additional feature, we also include the average number of passive constructions in a text. The parser underlying these features was the Alpino parser (van Noord et al. 2013), a state-of-the-art dependency parser for Dutch.

Semantic features

Since connectives serve as an important indication of textual cohesion in a text (Halliday and Hasan 1976, Graesser et al. 2004), we integrated several features based on a list lookup of connectives. These were drawn up by linguistic experts. As features, we counted the average number of connectives within a text and the average number of causal, temporal, additive, contrastive and concessive connectives on both the sentence and document level.

As named entity information provides us with a good estimation of the amount of world knowledge required to read and understand a particular text, we calculated the number of entities and unique entities, the number of entities on the sentence level and we made a comparison between predicted named entities – i.e. recognized by a NER system – and shallow entities – based on PoS-tags. To this purpose we used the NERD system (Desmet and Hoste 2013).

		FEATURE	FEATURE GROUP
Traditional	length-based (4)	average word length	*tradlen*
		average sentence length	
		ratio long words	
		% of polysyllable words	
	lexical (2)	% in frequency list	*tradlex*
		type token ratio	
Lexical	LM-based (2)	perplexity	*lexlm*
		normalized perplexity	
	terminology-based (2)	TF-IDF	*lexterm*
		Log Likelihood	
Syntactic	PoS-based (27)	average content words	*shallowsynt*
		average function words	
		average nouns	
		average type nouns	
		average nouns/sentence	
		average type nouns/sentence	
		average noun types	
		average adjectives	
		average type adjective	
		average adjective/sentence	
		average type adjectives/sentence	
		average adjective types	
		average verb	
		average type verb	
		average verb/sentence	
		average type verb/sentence	
		average verb types	
		average adverb	
		average type adverb	
		average adverb/sentence	
		average type adverb/sentence	
		average adverb types	
		average prepositions	
		average type prepositions	
		average prepositions/sentence	
		average type preposition/sentence	
		average preposition types	
	Dependency-based (6)	depth of the parse tree	*deepsynt*
		average sbars	
		average noun phrases	
		average verb phrases	
		average prepositional phrases	
		average passives	
Semantic	connectives-based (12)	average connectives/document	*shallowsem*
		average connectives/sentence	
		average causal/document	
		average causal/sentence	
		average temporals/document	
		average temporals/sentence	
		average additives/document	
		average additives/sentence	
		average contestive/document	
		average contestive/sentence	
		average concessives/document	
		average concessives/sentence	
	named entity-based (7)	number of entities	*semner*
		number of uniq entities	
		number of entities/sentence	
		number of uniq entities/sentence	
		number of ne/sentences	
		perc of ne	
		perc of regular entities	

Table 5.4: Exhaustive overview of all additional features that were implemented to perform readability prediction.

5.2 Exploring the added value of coreference and semantic role features

For all our experiments, we considered the readability prediction task as a classification task of text pairs. Two subtasks are defined:

1. **Binary classification**: determine for a given text pair whether text *a* is easier or more difficult than text *b*
2. **Multiclass classification**: here, multiple classes have to be predicted representing the five fine-grained readability levels between two texts, namely *left much easier – left slightly easier – equally difficult – right slightly easier – right much easier*

Considering the combined dataset we have at hand (Section 4.2.3), the 10,920 Dutch text pairs can be used as such for the multiclass classification task. For the binary classification experiments, we excluded all equally difficult pairs and put together the much and slightly easier or more difficult text pairs, leading to a reduced dataset of 6,084 Dutch text pairs. Both datasets are presented in Table 5.5

Subtask	**Acronym**	**Meaning**	Number of pairs
Binary	*LE*	left text easier	3042
	RE	right text easier	3042
Multiclass	*LME*	left text much easier	260
	LSE	left text somewhat easier	2782
	ED	both texts equally difficult	4836
	RSE	right text somewhat easier	2782
	RME	right text much easier	260

Table 5.5: Text pair statistics for both the binary and multiclass dataset.

When it comes to selecting the best features for readability prediction, there seems to be a consensus in readability research that first the correlation between the features and human assessments is measured (Pitler and Nenkova 2008, François 2011). The next step, if included at all, is then to see which features come out as good predictors when performing machine learning experiments such as regression (Pitler and Nenkova 2008), or classification (Feng et al. 2010) by in- or excluding features or feature groups from the prediction task.

Since we already derived our features based on the comments we received from our expert assessors, we start by performing the latter type of experiments to quickly move on to employing genetic algorithms to pinpoint the optimal feature combinations, something which has not been done before in readability prediction. We focus on feature group and individual feature selection for our two groups under consideration. Since each machine learning algorithm's performance is inherently dependent on the different parameters that are used, we perform joint optimization in both setups.

5.2.1 Machine learning method

For all classification experiments we employed Support Vector Machines, as preliminary experiments revealed this as the best learner for the tasks at hand. A Support Vector Machine (SVM) is a learning classifier capable of binary classification. It learns from the training instances by mapping them to a highdimensional feature space, and constructing a hyperplane along which they can be separated into the two classes. New instances are classified by mapping them to the feature space and assigning a label depending on their position with respect to the hyperplane. SVMs are said to have a robust generalization ability (Vapnik and Cortes 1995). For multiclass classification problems, separate SVMs have to be built. With the pairwise approach, one SVM is trained for every pair of classes. Another method is one versus the rest, where one SVM is built for each class to distinguish it from all other classes.

The SVM implementation used in our experiments is LibSVM[3], version 3.17 (Chang and Lin 2011). For our multiclass classification task we use the default paradigm which is pairwise multiclass classification.

With SVMs a lot depends on which kernel you decide to use to weigh the training instances in the new features space (see Cristianini and Shawe-Taylor (2000) for an in-depth discussion). With LibSVM four different kernels can be used: the Gaussian radial basis function (RBF) is the default option. Besides this RBF, a linear, polynomial or sigmoid kernel can also be chosen. For the linear kernel no additional kernel-specific parameters have to be set, the ones that have to be set for the other three kernel functions are summarized in Table 5.6 together with how they were configured for our purposes.

[3] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

	RBF	polynomial	sigmoid
Function	exp(-$\gamma \vert\vert x_i - x_j \vert\vert^2$)	($\gamma {x_i}^T x_j$+c^d)	tanh($\gamma {x_i}^T x_j$+c)
Parameters	free parameter γ: vary between 2^{-14} and 2^4, stepping by factor 4 (*default = 3*)		
		d: vary between 2 and 5 (*default = 1/number of features*)	
		c (constant trading off): fix to *default of 0*	

Table 5.6: Hyperparameters for the RBF, polynomial and sigmoid kernels.

Besides these kernel-specific settings, we configure the other hyperparameters as follows:

- We use the soft margin method to allow training errors when constructing the decision boundary, and vary the associated cost parameter C between 2^{-6} and 2^{12}, stepping by a factor of 4 (*default = 1*).
- Shrinking heuristics are always used, which is also the *default* option. Shrinking is a technique to reduce the training time: by identifying and removing some bounded elements in the optimization problem, it becomes smaller and can be solved in less time.
- The stopping criterion or ϵ is set to the *default of 0.001.* Because the optimization method only asymptotically approaches an optimum, it is terminated after satisfying this stopping condition.

Following the best practices as described in Hsu et al. (2003), we first scaled our feature set which means that all our features were linearly mapped to the range [0,1]. With unscaled features, values in greater numeric ranges dominate those in smaller numeric ranges. Another advantage of scaling is that it prevents numerical problems during the SVM calculation.

5.2.2 Evaluation metric

The choice of evaluation metric should be motivated by the task it is used for. We evaluate in terms of accuracy (Equation 5.8) since both our tasks are true classification tasks and because our datasets are not skewed in any way. As a consequence, we can value all classes as equally important.

$$accuracy = \frac{\text{true positives} + \text{true negatives}}{\text{total number of instances}} \tag{5.8}$$

Another important consideration when assessing the performance is to see whether the validation is carried out on a representative test set. We opted for 10-fold cross-validation[4].

As a baseline we calculate the majority class. For our binary classification task there is an equal distribution between our easier and harder class which results in a baseline of 50%. For the 5-way classification task, the equally difficult pairs

[4] The principle behind this was explained in Section 2.3.2.

form the majority class (4836 out of the 10920 possible text pairs), resulting in a majority baseline accuracy of 44.29%.

The main objective of our experiments is to find out whether coreference and semantic role features help for readability prediction. We hypothesize that adding this kind of information to a machine learning system should help, especially when this information is gold standard. Moreover, we believe that the added value of coreference and semantic roles features will be more outspoken for the multiclass classification task since this requires to discern more subtle differences between texts. We will now describe in close detail how we investigated this by conducting two different rounds of experiments.

5.2.3 Round 1

In a first round of experiments we wish to empirically verify whether our two feature groups under consideration actually contribute to the two classification tasks (binary and multiclass). To this purpose we manually exclude and include our two feature groups under consideration which results in the following four experiments:

(i) use all available features, except for the coreference and semantic role features;

(ii) use all available features, except for the semantic role features;

(iii) use all available features, except for the coreference features;

(iv) use all available features.

All these experiments were performed using the *default* hyperparameter settings of LibSVM and for each experiment we distinguish between a setup with the fully-automatic features and one with the gold standard ones.

5.2.4 Round 2

For the second round of experiments we are mainly interested whether, and if so how much, our coreferential and semantic role features contribute to the overall classification performance. Ideally, a feature vector should only contain highly informative features. We aim to test combinations of features rather than features in isolation in order to also take into account feature interaction and reduncancies.

A possible approach would be to perform a so-called hillclimbing procedure (Hoste 2005). Here, the starting point is an empty, complete or random feature set. In a next step, all neighbours of this current state are considered by either adding features – *forward selection* – removing them – *backward elimination* – or by using a combination of both techniques – *bidirectional hillclimbing*. Then, the feature set leading to the largest increase in performance is chosen and taken as new starting point. These two steps are repeated until no further improvement is obtained, the procedure then returns the final feature set as optimal feature set. The main problem with this approach is that it does not guarantee that the best solution is found and that the search algorithm converges to a local optimum (Hoste 2005). Contrary to hillclimbing approaches, genetic algorithms can allow multiple optima.

Besides feature selection, changing the hyperparameters of an algorithm can also have a dramatic effect on classifier performance (Hoste 2005, Desmet 2014). Most machine learning algorithms are configured to use sensible hyperparameters by default. These settings, however, are not guaranteed to be optimal for a particular problem. Previous research revealed that it is important to investigate the interaction between the parameters and feature representation (Mooney 1996, Hoste 2005). We hope that jointly optimizing both hyperparameters and features using genetic algorithms will allow for reliable results and comparisons.

We proceed to an optimization step by performing joint feature selection and parameter optimization. Essentially, this is an optimization problem which involves searching the space of all possible feature subsets and parameter settings to identify the combination that is optimal or near-optimal. Due to the combinatorially explosive nature of this type of experiment, a computationally feasible way of optimization had to be found. We decided to opt for a wrapper-based approach to feature selection using a genetic algorithm in conjunction with our learning method, LibSVM.

Genetic algorithms are an attractive approach to find optimal or near-optimal solutions. Standard references in GA-literature include Goldberg (1989) and Mitchell (1996). Genetic algorithms are search methods, based on the mechanics of natural selection and genetics. They require two things: Darwinian fitness-based selection and diversity. Figure 5.3 illustrates the basic principle behind GAs.

The procedure is rather simple: search starts from a population of individuals, who all present a candidate solution to the optimization problem to be solved. Applied to our dataset, the problem to be solved will be joint parameter optimization and feature selection. These individuals are typically represented as a bit string of fixed length, called a 'chromosome' or 'genome'. Each individual contains particular values for all algorithm parameters and for the selection of

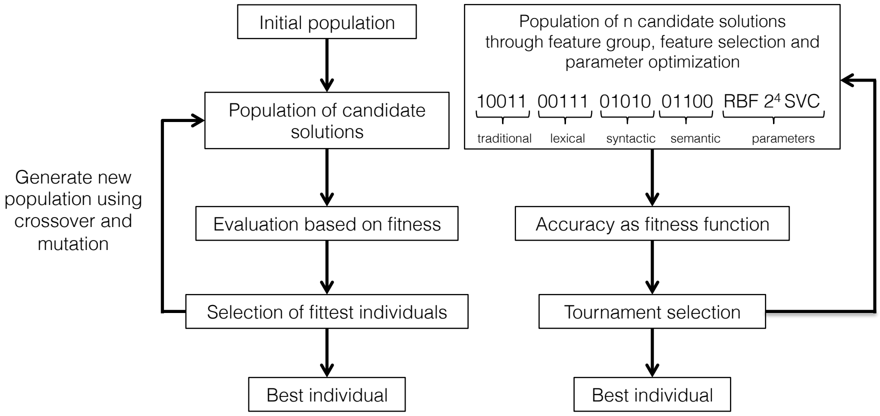

Initial population
Population of candidate solutions
Evaluation based on fitness
Selection of fittest individuals
Best individual
Generate new population using crossover and mutation
Population of n candidate solutions through feature group, feature selection and parameter optimization
10011 00111 01010 01100 RBF 2^4 SVC
traditional
lexical
syntactic
semantic
parameters
Accuracy as fitness function
Tournament selection
Best individual

Figure 5.3: Feature selection using a genetic algorithm approach. The left-hand side of the figure illustrates the general procedure, whereas the right-hand side translates this GA search to our task of readability prediction.

features. The population of chromosomes has a predefined size, larger population sizes increase the amount of variation present in the population at the expense of requiring more fitness function evaluations.

To decide which individuals will survive into the next generation, a selection criterion is applied defining how good the individual is at solving the problem, its fitness. For our experiments we run tenfold cross-validation on the training data and use the resulting classification accuracy value (Section 5.2.2) as the fitness score to be optimized.

After the fitness assignment, a selection method determines which individuals in the parent generation will survive and produce offspring for the next generation. We used the common technique of tournament-based selection (Goldberg and Deb 1991). Here, a fixed number of individuals is randomly picked from the population to compete in a tournament, where an individual's probability of winning is proportionate to its fitness. The winner is selected as parent. This process is repeated as many times as there are individuals to be selected. Unless the stopping criterion is reached at an earlier stage, optimization stops after a predefined set of generations.

We performed joint optimization in two different setups. In both setups we allow 100 generations and set the stopping criterion to a best fitness score that remained the same during the last five generations.

1. We perform hyperparameter and feature group selection using the ten feature groups we have available (i.e. tradlen, tradlex, lexlm, lexterm, shallowsynt, deepsynt, shallowsem, semner, coref and srl). We allow variation in LibSVM's hyperparameters as described in Section 5.2.1. Here, our search starts from a population of 100 individuals.
2. We perform hyperparameter selection and at the same time freeze the features within the groups that are not our prime focus (i.e. these features are all turned on) and allow individual feature selection among the features derived from coreference (5 features) and semantic role information (20 features). We allow the same hyperparameter variation. Here, our search starts from a population of 300 individuals to ensure sufficient variation.

We performed all optimization experiments using the Gallop toolbox (Desmet et al. 2013). Gallop provides the functionality to wrap a complex optimization problem as a genome and to distribute the computational load of the GA run over multiple processors or to a computing cluster. It is specifically aimed at problems involving natural language.

For our experiments we ran Gallop on a high performance cluster consisting of many worker nodes, as visualized in Figure 5.4. The left-hand side of the figure represents the calling script that can be used to set the GA (everything that was explained in Figure 5.3). The right-hand side illustrates how the GA manager within Gallop submits each generation as an array of job requests to be processed simultaneously, i.e. it polls the cluster until all jobs are finished[5].

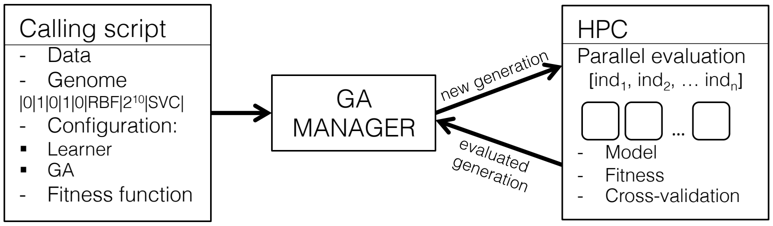


Figure 5.4: Running Gallop on a high performance cluster.

[5] This work was carried out using the STEVIN Supercomputer Infrastructure at Ghent University, funded by Ghent University, the Flemish Supercomputer Center (VSC), the Hercules Foundation and the Flemish Government, department EWI.

5.2.5 Summary

We can summarize the entire experimental setup as follows:

- Information sources
    - 87 individual features which can be categorized in 10 feature groups
    - Two different ways of deriving coreference and semantic role features
        - Fully-automatic
        - Based on gold standard information
- Classification tasks
    - Binary classification
    - Five-way or multiclass classification
- Machine learner: LibSVM
- Experiments to explore the added value of coreference and semantic role features:
    - Round 1
        - − coref − srl
        - + coref − srl
        - − coref + srl
        - + coref + srl
    - Round 2
        - Hyperparameter and feature group selection
        - Hyperparameter and individual feature selection

CHAPTER 6

Results and discussion

In this chapter, we present the results of our two rounds of readability prediction experiments. The first round consisted of experiments where the default hyperparameter settings of LibSVM were used and where the two semantic feature groups were manually added or excluded. In the second round of experiments, we used a GA-based search to determine the optimal feature groups and individual feature combinations, while at the same time optimizing the LibSVM hyperparameters.

The results are presented in Section 6.1. We start with two baselines: a majority baseline and a first experiment where the default LibSVM settings are used and no coreference or semantic role features are added as features. In a next step, we explore the added value of using automatically predicted coreference and semantic role features in both of our rounds, to end with the results of using gold-standard semantic features instead of automatically predicted ones. In Section 6.2, we discuss and analyze in close detail the added value of our two semantic layers.

6.1 Results

6.1.1 Baselines

In Table 6.1, we present the results of the majority class and the first two experiments where all features except the coreference and semantic role features were included and the default hyperparameter settings for LibSVM were set.

	BINARY	MULTICLASS
Majority class	50.00	44.29
All features (– coref – srl)	91.16	59.01

Table 6.1: Baseline results expressed in accuracy.

In general, we see that for both tasks we are able to beat the majority baselines. As expected, the performance on our binary dataset is much higher than on the multiclass dataset. In the further discussion of the experimental results, we consider these two results, a classification accuracy of 91.16% for the binary and one of 59.01% for the multiclass task, as the baselines against which to compare the results when coreference and semantic role features are included.

6.1.2 Adding coreference and semantic roles

Table 6.2 present the results when including coreference and semantic role features, which were derived from the output of the COREA and SSRL systems. The baseline results of the first experiment are added for comparison (– coref, – srl). In the upper part, we present the experiments where the features were manually in- or excluded from our system using the default parameters. The lower part of the table contains the results of the wrapper-based optimization experiments. The results of these two different rounds will be discussed separately.

Round 1: We observe some differences between both classification tasks. In the binary classification task, it seems as though only the semantic role features account for a higher performance; when adding coreference features we are not able to beat the baseline (accuracy of only 91.12% versus one of 91.16%). With the addition of the semantic roles alone we achieve the highest performance (93.28%), if both semantic layers are added, however, the score decreases again (93.08%).

For the multiclass experiments, we observe that the overall performance gain

		BINARY	MULTICLASS
Round 1	– coref – srl	91.16	59.01
	+ coref – srl	91.12	59.28
	– coref + srl	**93.28**	59.44
	+ coref + srl	93.08	**59.49**
Round 2	Joint feature groups	98.01	73.12
	Joint individual features	**98.18**	**73.73**

Table 6.2: Accuracy of the classifications when adding automatically-derived coreference and semantic role features in both rounds of experiments.

is much lower when this type of semantic information is added to our feature vectors (the baseline is 59.01% whereas the best result we achieve is 59.49%). Again, the best improvement over the baseline is achieved when semantic role features are included, but for the multiclass experiments coreference also seems to contribute. This improvement, however, is rather marginal (from an accuracy of 59.44 with only semantic roles to one of 59.49 with both information sources). This brings us to the results of the feature selection experiments using genetic algorithms.

Round 2: Overall, we observe that both tasks benefit a lot from jointly optimizing the hyperparameters and features. From a best score of 93.28% using the default LibSVM hyperparameters and all available features except the coreference features to an optimal score of 98.19% where both the hyperparameters and all semantic features are optimized for the binary classification task. And from an accuracy of 59.49% using the default parameters and all available features to an optimal score of 73.73% where both the hyperparameters and semantic features are optimized for the multiclass classification task.

If we compare both optimization setups, viz. jointly optimizing feature groups versus jointly optimizing individual semantic features, we observe that the best results are achieved with the individual feature selection experiments. The differences in performance between both optimization rounds, however, are not that outspoken: a difference of 0.17 points for the binary and one of 0.61 points for the multiclass classification task.

Overall, from the experiments of *Round 1*, we can conclude that adding semantic information in the form of automatically-derived coreference and semantic role features seems to improve the performance of our readability classifier for both classification tasks. More information about which hyperparameters and features were selected in *Round 2* will be given in Section 6.2.

6.1.3 Using gold standard semantic information

In Table 6.3 we present the results of the experiments using features which were derived from gold-standard annotations (Gold). These results are directly compared to the results we achieved in the previous section (Auto), indicated in gray. The results that are bold-faced are the best overall results we achieved for the two classification tasks.

		BINARY		MULTICLASS	
		Gold	*Auto*	*Gold*	*Auto*
Round 1	– coref – srl	91.16	91.16	59.01	59.01
	+ coref – srl	91.26	91.12	59.78	59.28
	– coref + srl	92.23	**93.28**	60.05	59.44
	+ coref + srl	92.66	93.08	**61.09**	59.49
Round 2	Joint feature groups	97.50	**98.01**	71.76	**73.39**
	Joint individual features	97.62	**98.19**	72.44	**73.73**

Table 6.3: Results of using gold-standard coreference and semantic role features in both rounds of experiments.

What immediately draws the attention is that in almost all cases the best results are achieved when using the automatically predicted features. This is contrary to our intuition, as we hypothesized that the best results would be achieved when using gold standard information. We will now discuss the results of both rounds of experiments in closer detail, our focus is on the gold-standard setup.

Round 1: We notice that the semantic roles seem to contribute more than the coreference features for both tasks. In the binary task, a difference with the fully-automatic setup is that now the coreference features alone also constitute a slightly better performance, an accuracy of 91.26% versus the baseline of 91.16%. The best result using gold standard information, however, is not able to beat the best result achieved in the fully-automatic setup: an accuracy of 92.66% versus one of 93.28%. In the multiclass setup, on the contrary, we do observe that the best performance is achieved when the semantic features are derived from gold standard information. The accuracy is always higher in the golden setup and the best result is achieved using both gold standard coreference and semantic role information, an accuracy of 61.09%.

Round 2: Studying the optimization experiments using gold features, we notice that, for both tasks, the best results are achieved when performing joint hyperparameter and individual semantic feature selection. The results of the automatic setup, however, are never outperformed by our gold-standard setup: an accuracy of 97.62% versus one of 98.19% for the binary task and one of

72.44% versus 73.73% for the multiclass classification task.

We can conclude that the upper bound of our readability prediction system using the information sources we have available is reached when using automatically predicted semantic information. We will now continue to a more in-depth analysis and discussion of these results in order to shed more light on this matter. But first we investigate whether these results are statistically significant.

6.1.4 Statistical significance of the results

Judging from the raw performance results, coreference and semantic role information helps classifying readability when added as additional features to our feature vectors. In a next step, we investigated whether these results are statistically significant.

To this purpose, we applied the bootstrap resampling test (Noreen 1989). Bootstrap samples (n=5000) were randomly drawn with replacement from the output of each (optimal) system, i.e. from the set of classified instances. This was done 3000 times and classification accuracy was calculated on every sample. We report the mean accuracy and standard error over all samples. If the standard error band of a system does not overlap with the band of another system, their accuracies can be said to differ significantly, with a confidence interval of > 0.95.

These results are presented in Figure 6.1. Considering the *Round 1* experiments, each time represented on the left-hand side of the plots and situated lower than the optimized results. For the binary classification task, we observe that when we incorporate these features in their automatically-derived form, only the semantic roles seem to truly contribute and lead to a statistically significant improvement. For the gold-standard setup the results are only borderline statistically significant when both the semantic roles and coreference information are included as features. For the multiclass task, we observe that none of the performance increases obtained by incorporating semantic information, even not in the best gold setting, are statistically significant.

In all plots we observe that the optimized results (*Round 2*) far outperform the results of manually in- or excluding these features using the default LibSVM settings (*Round 1*). The difference between both rounds is each time clearly statistically significant. If we have a closer look at the two experiments that were performed in the *Round 2* setting, we observe that the difference between optimizing on the feature groups or allowing individual feature selection, the latter step of which always led to the top performance, is not statistically significant.

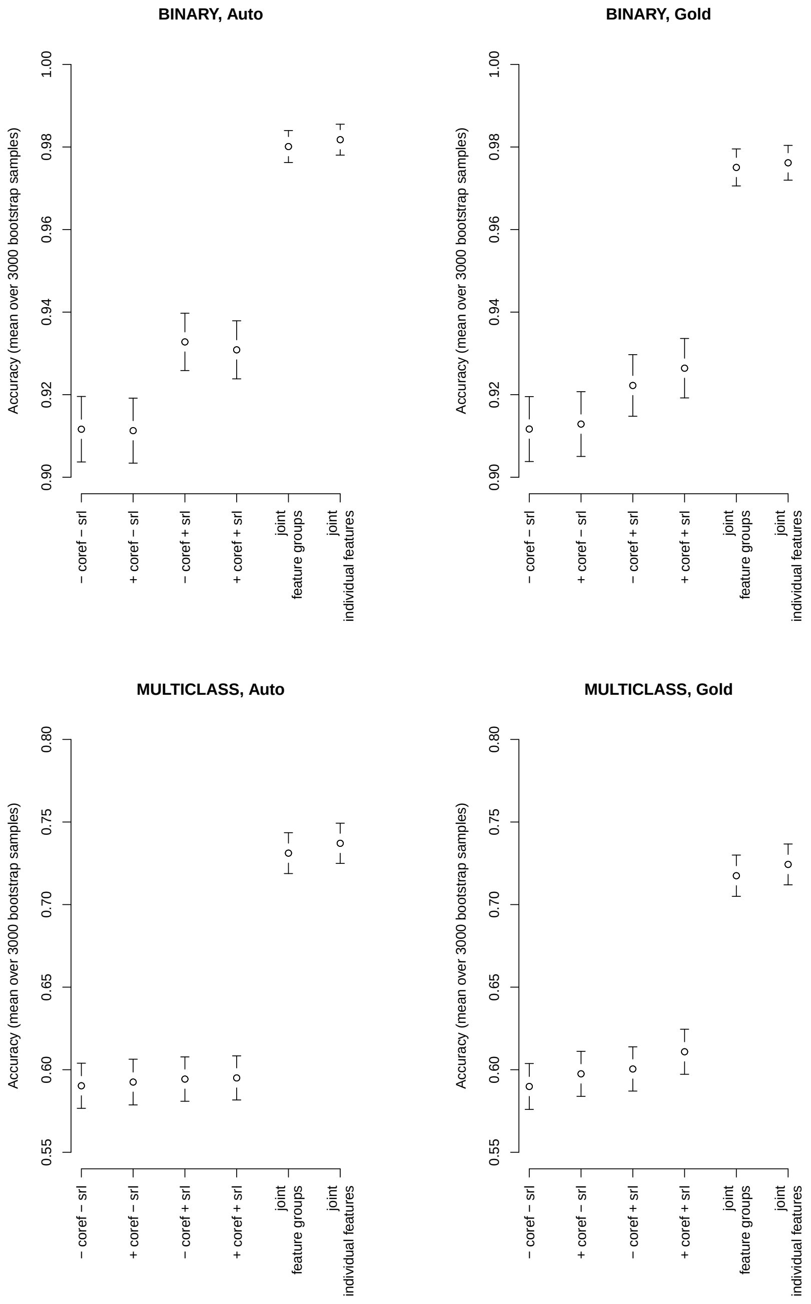


Figure 6.1: Plots representing the system accuracy and standard error of the bootstrap resampling tests that were performed for all readability experiments.

6.2 Discussion

We will start our discussion with a qualitative error analysis of the default, viz. *Round 1*, experiments (Section 6.2.1). Here, we observed that we always got better results when including semantic roles and to a lesser extent when including coreference.

In a next step we will discuss in closer detail which hyperparameters, feature groups and individual semantic features came out of our joint optimization, viz. *Round 2*, experiments (Section 6.2.2). This should help us understand whether the coreference and semantic role features actually contributed to the top performance or whether these improvements were merely a result of the hyperparameter optimization.

6.2.1 Error analysis

In our first round of experiments, we manually in- or excluded coreference and semantic role features using the default LibSVM parameters. We manually inspected the output of both classification tasks, focussing on which text pairs were classified correctly in the most successful setup when compared to the baseline experiments (– *coref,* – *srl*, accuracy of 91.16%). We will present a number of examples, which are all texts derived from our readability corpus and that were combined into text pairs for our classification tasks.[1]

Given the high number of features and given the fact that our feature values were almost always averaged over text length (cfr. Section 5.1), it is difficult to pinpoint which features or feature groups constitute an added value. Nevertheless, we will make such an endeavour for our two semantic layers. We can state with certainty that all the text pairs presented next were only classified correctly after features derived from these semantic layers were added.

For the **binary** classification experiments, the best results in the automatic setup were achieved with only the semantic roles (– *coref* + *yes srl*, accuracy of 93.28%) and with both the coreference and semantic roles in the gold-standard setup (+ *coref* + *srl*, accuracy of 92.66%).

In the binary experiments, the text pairs 19–20, 21–22 and 23–24 were all classified incorrectly when semantic information was excluded from the instancebase.

(19) De Israëlische premier Ariel Sharon verhindert de Palestijnse president Yasser Arafat de kerstnachtmis bij te wonen in Bethlehem. Sharon wou

[1] The English translation of these examples can be found in Appendix B.

Arafat alleen toestemming geven als hij de daders van de moord op de Israëlische minister van Toerisme Revahan Zeevi oppakte. Het is de eerste keer dat Arafat de mis niet kon bijwonen sinds Bethlehem in 1995 overgedragen werd aan de Palestijnse Autoriteit. Het verbod lokt kritiek uit. Onder meer paus Johannes Paulus II, de Europese Unie en de Verenigde Naties veroordelen de maatregel. Een hoge functionaris van de radicale islamitische organisatie Jihad kondigt aan te stoppen met aanslagen tegen Israël. Jihad wil naar eigen zeggen de eenheid onder de Palestijnen bewaren.
• *coreference*: seven coreference chains are found in this text, all smaller ones.
• *semantic roles*: on average four PropBank roles, mostly arguments, are activated per sentence.

(20) Twee tienermeisjes van 13 en 15 jaar oud springen samen van de tiende verdieping van een Brussels appartementsgebouw. De meisjes verbleven allebei in een psychiatrische instelling, het jongste vanwege depressie. Ze brachten een ongeoorloofd bezoek aan de vader van een van beiden. Toen die van zijn boodschappen terugkeerde, deed hij de macabere ontdekking. Volgens een onderzoek van de Franstalige Brusselse universiteit ULB uit 1999 zijn er in Brussel jaarlijks zo'n 800 tot 1100 zelfmoordpogingen bij jongeren tussen de 12 en de 19 jaar. In heel België is zelfmoord al jaren een van de voornaamste doodsoorzaken bij tieners.
• *coreference*: contains three chains, two of these comprise up to four coreferential relations.
• *semantic roles*: on average, 2.13 PropBank roles per sentence.

Without semantic information, the text pair consisting of examples 19 and 20 was classified as '-1' which means that example 19 was considered easier than example 20, whereas it ought to be classified as '1', implying the opposite. In the best setups, we observed that both the fully-automatic and gold-standard setup were able to correctly classify this text pair. Since the fully-automatic setup only considered the semantic role features, we could claim that especially the semantic roles helped to correctly classify this text. If we have a closer look at examples 19 and 20 we do observe that example 19 contains more complex sentences with especially a higher number of arguments.

(21) Het Zimbabwaanse parlement keurt twee omstreden wetten goed die de regering verregaande macht toekennen. Volgens de populaire oppositiepartij MCD wil president Robert Mugabe hen met die wetten uitschakelen in de aanloop naar de presidentiële verkiezingen in maart. De nieuwe wetten maken onder meer kritiek op de president illegaal. De internationale gemeenschap bekritiseert de wetten en het beleid in

Zimbabwe. Vooral het drieste verdrijven van blanke boeren bij de landhervorming en het beknotten van de persvrijheid leveren negatieve reacties op. Vrijdag 11 januari verschijnt een Zimbabwaanse delegatie voor Europese diplomaten in Brussel, die hen ondervragen over vermeende schendingen van de mensenrechten. Europa denkt erover de samenwerking en geldelijke steun van 200 miljoen euro stop te zetten als de situatie niet verbetert. Verder eist Europa dat het land internationale waarnemers en pers toelaat tijdens de verkiezingen in maart.
• *coreference*: this text contains six chains, mostly small ones switching from one sentence to the other.
• *semantic roles*: on average three PropBank roles are evoked per sentence.

(22) Zie hier de elementen voor het antwoord op uw vragen. De meegedeelde gegevens zijn niet gestandaardiseerd en houden geen rekening met de eventueel verschillende samenstelling van de leeftijdsklassen en het geslacht van de betrokken populaties. 1) In Vlaanderen zijn in 20068.349 appendicectomieën verricht voor een totaal bedrag van 1.653.722 euro. 2) In Wallonië zijn in 2006 4.527 appendicectomieën verricht voor een totaal bedrag van 902.807 euro. 3) Op basis van de bevolkingsgegevens 2006 die beschikbaar zijn op de FOD Economie, KMO, Middenstand en Energie, zijn er in Vlaanderen 137 appendicectomieën per 100.000 inwoners, tegenover 132 in Wallonië. 4) De kosten voor de analyses inzake klinische biologie voor de ziekenhuisopnames in het raam van een appendicectomie bedragen 16,14 euro in Vlaanderen, tegenover 23,67 euro in Wallonië.
• *coreference*: contains two smaller chains consisting of two and three noun phrases, respectively.
• *semantic roles*: on average, 3.7 PropBank roles are found, mostly modifiers.

Considering the text pair consisting of examples 21 and 22 we observed that only the setup using gold-standard features was able to correctly classify this text pair as '-1' which implies that example 21 is easier than example 22. Since the gold-standard setup considers both coreference and semantic roles features, we could say that the coreference features provided the tipping point for this classification. We do see that in example 21 more coreferential links are made between the different sentences, six chains, whereas example 22 is merely a listing of various items and contains only two small chains.

(23) Op het Europese congres voor urologie dat van 16 tot 19 maart 2005 heeft plaatsgevonden in Istanboel, werd een zitting gewijd aan de behandeling van stressincontinentie door plaatsing van een suburetraal bandje.

Dankzij de kwaliteit van de textuur van het bandje kent die chirurgische techniek meer en meer opgang in de medische wereld. De ingreep die eerst langs retropubische weg wordt uitgevoerd (TVT), geeft uitstekende resultaten: ongeveer 90% van de patiënten wordt weer continent. Er werd een nieuwe techniek gepresenteerd waarbij het bandje door het foramen obturatorium (TOT, trans obturator tape) wordt gevoerd. Die techniek is eenvoudig en riskeert niet de blaas te verwonden. Bij die techniek dient niet eerst een cystoscopie te worden verricht. Tijdens deze sessie hebben meerdere groepen de 2 technieken vergeleken. Uit de verschillende presentaties blijkt dat beide chirurgische technieken als weinig invasief kunnen worden beschouwd.
• *coreference*: contains five chains, three short ones and two of moderate length.
• *semantic roles*: on average 4.5 semantic roles are found, equal subdivision between arguments and modifiers.

(24) Lambermont wordt dikwijls gebruikt als aanduiding van de ambtswoning van de Eerste Minister in België (een beetje zoals Downingstreet in Londen). Het is gelegen aan de hoek van de Lambermontstraat en de Hertogstraat te Brussel, niet ver van het Koninklijk Paleis en het Paleis der Natiën. Deze ambtswoning is niet te verwarren met de Wetstraat 16, waar het Kabinet van de Eerste Minister is gevestigd. Het gebouw is genoemd naar baron Auguste Lambermont (1819-1905), die als hardwerkende secretaris-generaal van het Ministerie van Buitenlandse Zaken de Belgische diplomatie vanuit Brussel gedurende de hele negentiende eeuw domineerde. Als groot bepleiter van het economisch liberalisme was hij n van de stuwende krachten om door middel van douaneunies de Belgische markt te openen voor haar buurlanden. Hij speelde onder andere een cruciale rol bij de afkoop van de Scheldetol van Nederland (1863) en de koloniale avonturen van Leopold II. Zijn onderhandelingskunst en diplomatiek inzicht bleken beslissend tijdens de verschillende conferenties, die de erkenning van de Onafhankelijke Congostaat als persoonlijke privétuin van Leopold II bewerkstelligden. Het gebouw gaf ook zijn naam aan het Lambermontakkoord de staatshervorming van 2001.
• *coreference*: contains eight coreference chains, five of which consist of only two elements and two rather long chains (i.e. Lambermont referring to the building and Lambermont referring to the person).
• *semantic roles*: contains on average 3.75 semantic roles per sentence.

Finally, the text pair with examples 23 and 24 in it was only classified correctly using the fully-automatic setup. The correct classification is ‘1’ implying that example 24 is easier than example 23. Again, the semantic roles made the final call here, which is clearly visible in example 23 which contains more complex sentences (4.5 semantic roles on average).

For the **multiclass** classification experiments, the best results were achieved in both the automatic and gold-standard setup when both the coreference and semantic role features were included (*+ coref + srl*, accuracy of 61.09% in the gold-standard and one of 59.49% in the fully-automatic setup).

We performed a similar analysis of the multiclass experiments' output. In these experiments the text pairs 24–25, 26–27 and 20–28 were all classified incorrectly when excluding semantic information from our instancebase.

(25) De senaatscommissie Binnenlandse Zaken rondt het algemene debat af over het stemrecht voor ingeweken burgers die niet uit de Europese Unie komen. Enkele socialistische senatoren halen fel uit naar VLD-voorzitter Karel De Gucht, die enkele dagen eerder in de kranten een absoluut njet uitsprak. De vijf andere coalitiepartners zien dat als een dictaat. De Vlaamse liberalen willen vooral vermijden dat het tot een stemming komt over het voorstel. Gezien de huidige verhoudingen zouden acht van de vijftien commissieleden op het groene knopje duwen. Bij monde van senator Jeannine Leduc laat de VLD voor het eerst openlijk blijken dat ze het eigen voorstel voor een spijtoptantenregeling heeft bevroren opdat de PS het migrantenstemrecht niet zou goedkeuren. Leduc herinnert de Waalse socialisten aan die afspraak. Na afloop spreekt Philippe Moureaux (PS) al verzoenende taal. Hij wil de kwestie eerst uitklaren binnen de meerderheid.
• *coreference*: contains seven coreferential chains, all except two are rather short.
• *semantic roles*: on average 4.89 PropBank roles, especially arguments, are activated.

The text pair consisting of examples 24 and 25 should be classified as '0' which means that both texts can be considered equally difficult. The default setup with no coreference or semantic role information classified this text pair as 50, which means that it considered example 24 as more difficult than example 25. Both the gold-standard and automatic setup correctly classified this text pair afterwards. If we consider both text 24 and 25 we see that these texts do contain quite difficult material. They both contain a high number of short coreference chains and a high number of semantic roles are activated.

(26) De Verenigde Naties en de regering van Sierra Leone ondertekenen een akkoord over de oprichting van een oorlogstribunaal. Het hof zal zich uitspreken over verdachten van gruweldaden tijdens de burgeroorlog waarin

50.000 doden vielen. Vorige week werd het einde van die oorlog officieel afgekondigd. De strijd begon in 1991 met de acties van het Revolutionary United Front (RUF) onder leiding van Foday Sankoh, later mengden andere groeperingen zich in het conflict, ook buitenlandse. Het RUF werd berucht omdat het tegenstanders en burgers verminkte door armen af te hakken en doordat het vele kindsoldaten inlijfde. Het nieuwe tribunaal verschilt van de andere VN-tribunalen voor ex-Joegoslavië en Rwanda. Die laatste twee werden opgericht vanuit de VN, het nieuwe tribunaal komt er op verzoek van het land zelf en zal bestaan uit VN-rechters en rechters uit Sierra Leone zelf.
• *coreference*: contains seven coreferential chains, three of which span more than half of the text.
• *semantic roles*: contains on average 4.3 PropBank roles in each sentence.

(27) * U kunt de toon, het volume en de snelheid naar wens regelen door de toetsen omhoog of omlaag in te drukken. De toetsen voor toon, volume en snelheid bevinden zich midden boven op de bovenzijde, respectievelijk van links naar rechts. * Om snel terug of vooruit te spelen in een boek, houdt u de toets Vooruit of Terugspoelen ingedrukt tot u de gewenste positie in het boek hebt bereikt. Wanneer u de knop loslaat, gaat u automatisch weer terug naar de gewone afspeelsnelheid.
• *coreference*: comprises five chains, four of two elements and one of four elements.
• *semantic roles*: contains on average 6.5 PropBank roles per sentence (equal amount of modifiers and arguments).

The text pair consisting of examples 26 and 27 was classified as '0' or equally difficult in both the default and gold-standard setup when semantic information was excluded from the instancebase. In the follow-up experiments, only the fully-automatic setup correctly classified text 26 as easier than text 27. We observe that especially the average amount of activated PropBank roles is larger in text 27.

(28) Welke voordelen bleek Zyprexa Velotab tijdens de studies te hebben? Zyprexa Velotab was evenals Zyprexa werkzamer in termen van symptoomverbetering dan placebo (schijnbehandeling). Zyprexa Velotabtabletten waren even werkzaam als de vergelijkingsmiddelen voor de behandeling van schizofrenie, van matige tot ernstige manische episoden en ter voorkoming van een recidief bij patiënten met een bipolaire stoornis. Welke risico's houdt het gebruik van Zyprexa Velotab in? De meest voorkomende bijwerkingen van Zyprexa Velotab (waargenomen bij meer dan 1 op de 10 patiënten) zijn somnolentie (slaperigheid), gewichtstoe-

name en een verhoogde concentratie prolactine (een hormoon) in het bloed. Zie de bijsluiter voor de volledige beschrijving van de gerapporteerde bijwerkingen van Zyprexa Velotab. Zyprexa Velotab mag niet worden gebruikt bij mensen die mogelijk overgevoelig (allergisch) zijn voor olanzapine of voor enig ander bestanddeel van het middel. Zyprexa Velotab mag niet worden gebruikt bij patiënten met een verhoogd risico op nauwe-kamerhoekglaucoom (verhoogde druk in het oog).
• *coreference*: contains five coreference chains (one long one referring to zyprexa velotab) and four consisting of only two elements.
• *semantic roles*: on average 2.13 semantic roles are activated in each sentence.

Finally, the text pair comprising examples 20 and 28 was only able to be correctly classified in the gold-standard setup which recognized example 20 as easier than example 28. Intuitively, when reading text 28, we are surprised that the lexical features where not able to make this distinction since it clearly contains a more complex vocabulary. Considering the semantic roles, the number of activated PropBank labels per sentence is the same in both texts, which means the tipping point came from the coreference features. If we have a closer look, the only notable difference between these two examples is that text 28 contains less unique coreferential noun phrases because of the constant repetition of the noun phrase ‘zyprexa velotab’.

6.2.2 Analysis of the joint optimization experiments

Since the joint optimization experiments achieved the best results, we will now discuss which hyperparameters, feature groups and individual semantic features emerged from the fittest individuals.

Since at the end of a GA optimization run, the highest fitness score may be shared by multiple individuals that have different hyperparameters or features (Desmet 2014), we decided that runner-up individuals to that elite can also be considered valuable solutions to the search problem. To this purpose, we refer to the k-nearest fitness solution set; these are the individuals that obtained one of the top k fitness scores, given an arithmetic precision. Following Desmet (2014), we used a precision of four significant figures and set k to three.

We will in each case discuss which hyperparameters and especially which features groups and individual semantic features were selected. The features are visualized using a colour range: the closer to blue, the more this feature group or feature was turned on and the closer to red, the less important the feature group or feature was for reaching the optimal solution. The numbers within the cells represent the same information but then percentagewise.

Selected hyperparameters and feature groups

This setting included optimization runs where 100 individuals were assessed in one generation. The best results for both classification tasks were achieved in the setting with automatically-derived features for coreference and semantic roles, reaching an accuracy of 98.01% and 73.39%, respectively. For both tasks and setups, the number of generations ranged from eight to ten, which implies that the genetic algorithm quickly converged to an optimal solution. We will now explain in closer detail which hyperparameters were set and which feature groups were retained in the fittest individuals.

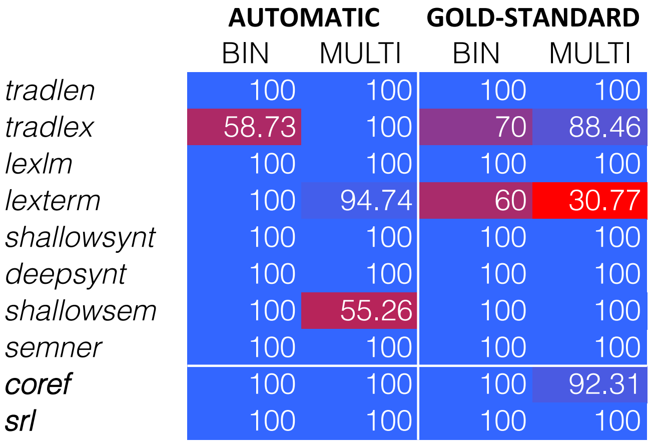
Table 6.4 gives an overview of the selected hyperparameters after each run.

	Default	**AUTOMATIC**		**GOLD-STANDARD**	
		BIN	MULTI	BIN	MULTI
Kernel	RBF	linear	RBF	linear	linear, RBF
Cost-value	1	2^{10}, 2^{12}	2^{12}	2^{12}	2^{10}, 2^{12}
γ	3	n/a	2^{-6}	n/a	0

Table 6.4: Selected hyperparameters for the joint optimization task with feature groups.

Machine learning algorithms are configured to use sensible hyperparameters by default. Our optimization experiments clearly show, however, that kernel preference is affected by the task. For the binary classification task, linear kernels are always selected, whereas for the multiclass classification, the default RBF kernels are favoured, indicating that a more complex kernel is beneficial for the multiclass task and that this task is more complex. For all setups, a high cost value C is selected (range between 2^{10} and 2^{12}) and the γ value for RBF kernels is very small or zero.

In Figure 6.2 we illustrate which feature groups were considered important using the above-mentioned colour range. Considering our two semantic groups – coref and srl – we see that the semantic roles are turned on all the time, regardless of whether these were derived automatically or from gold standard information. For coreference, the same seems to apply, except that in the multiclass setup using gold standard information the added value of the coreference features seems less outspoken, though this group was turned on in more than 9 of the 10 cases when the optimal solution was reached.



	AUTOMATIC		GOLD-STANDARD	
	BIN	MULTI	BIN	MULTI
tradlen	100	100	100	100
tradlex	58.73	100	70	88.46
lexlm	100	100	100	100
lexterm	100	94.74	60	30.77
shallowsynt	100	100	100	100
deepsynt	100	100	100	100
shallowsem	100	55.26	100	100
semner	100	100	100	100
coref	100	100	100	92.31
srl	100	100	100	100

Figure 6.2: Illustrating which feature groups were selected in the joint optimization experiments in all different setups (automatic, gold, binary and multiclass).

Looking at the other feature groups, we notice that five feature groups received the same status as the semantic roles group: the traditional length-related features (tradlen), the lexical features derived from language modelling (lexlm), the shallow syntactic (shallowsynt), the deep syntactic (deepsynt) and the named entity features. If we look at the remaining groups, we observe some differences between the binary and multiclass setup and between the gold-standard and fully-automatic setup, which indicates a differing interplay.

Contrary to previous research (Feng et al. 2010, Feng 2010), we observe that using our dataset and technique we do observe that both the coreference and semantic role features turn out to be important predictors for readability prediction.

Selected hyperparameters and individual features

This setting included optimization runs where 300 individuals were assessed in one generation. The best results for both classification tasks were achieved in the setting with automatically-derived features for coreference and semantic roles, reaching an accuracy of 98.19% and 73.73%, respectively. For the binary task, the number of generations ranged from 10 to 12 for both the automatic and gold-standard setup. For the multiclass task, the genetic algorithm search required more generations to converge: from 25 for the automatic to 34 for the gold-standard setup, indicating that the multiclass task is more sensitive to optimization. We will now again explain in more detail which hyperparameters were set and which individual coreference and semantic role features were retained in the fittest individuals.

Table 6.5 gives an overview of the selected hyperparameters after each run.

	Default	**AUTOMATIC**		**GOLD-STANDARD**	
		BIN	MULTI	BIN	MULTI
Kernel	RBF	linear	RBF	linear	RBF
Cost-value	1	2^{10}, 2^{12}	2^4, 2^8	2^{12}	2^{10}
γ	3	n/a	2^{-2}	n/a	2^{-6}

Table 6.5: Selected hyperparameters for the joint optimization task with individual feature selection.

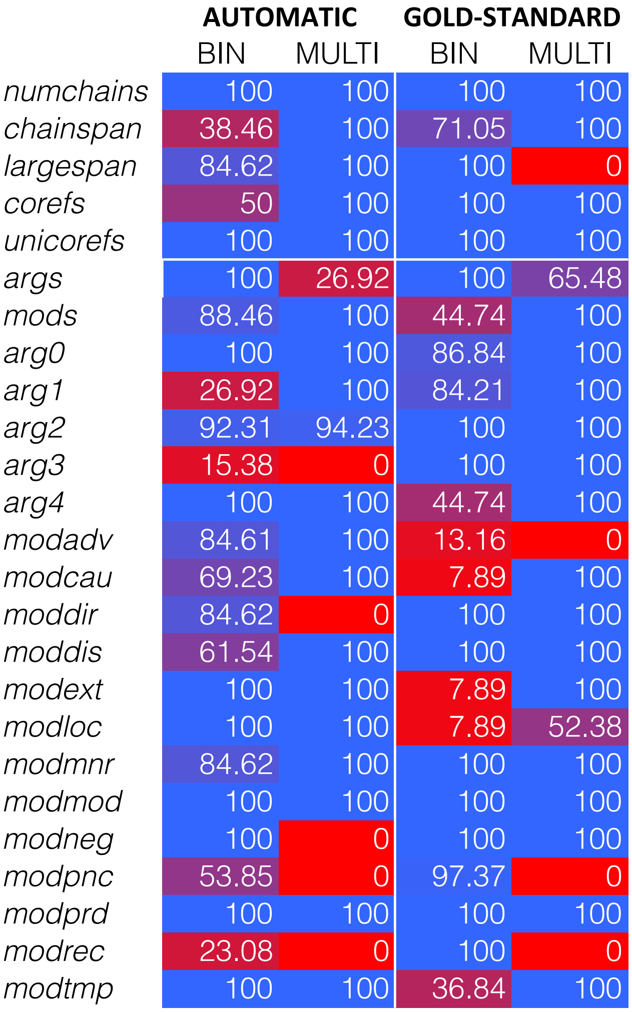
We observe more or less the same tendencies as in the previous optimization experiment: the same kernels are chosen for both tasks, a linear kernel for the binary and RBF for multiclass classification. The choice of C for the multiclass tasks is slightly lower, 2^4 to 2^{10} versus 2^{10} to 2^{12} for binary. Again, the value for γ is small.

Figure 6.3 illustrates which individual semantic features were retained in the fittest individuals and can thus be considered most important. The upper part of the figure presents the coreference features, whereas the lower part lists the semantic role features that were retained in all different setups.

Overall, we notice that more features are turned on in the multiclass classification tasks (18 blue cells in the automatic, and 19 in the gold-standard setup). Also, the percentages in the other cells are closer to 100 or 0.

This is interesting information, because it indicates that for the multiclass classification task, which is the more difficult task (five labels have to be predicted, overall lower results are achieved, the more complex RBF kernel is chosen), it is more beneficial to have more semantic information available in the feature vectors.

Looking at the two feature groups, we observe that, in both groups, only two features are always turned on regardless of the task or feature setup. For the coreference group that is the total number of chains (numchains) and the average number of unique coreferring NPs (unicorefs).



	AUTOMATIC		**GOLD-STANDARD**	
	BIN	MULTI	BIN	MULTI
numchains	100	100	100	100
chainspan	38.46	100	71.05	100
largespan	84.62	100	100	0
corefs	50	100	100	100
unicorefs	100	100	100	100
args	100	26.92	100	65.48
mods	88.46	100	44.74	100
arg0	100	100	86.84	100
arg1	26.92	100	84.21	100
arg2	92.31	94.23	100	100
arg3	15.38	0	100	100
arg4	100	100	44.74	100
modadv	84.61	100	13.16	0
modcau	69.23	100	7.89	100
moddir	84.62	0	100	100
moddis	61.54	100	100	100
modext	100	100	7.89	100
modloc	100	100	7.89	52.38
modmnr	84.62	100	100	100
modmod	100	100	100	100
modneg	100	0	100	100
modpnc	53.85	0	97.37	0
modprd	100	100	100	100
modrec	23.08	0	100	0
modtmp	100	100	36.84	100

Figure 6.3: Illustrating which individual coreference and semantic role features were selected in the joint optimization experiments in all different setups (automatic, gold, binary and multiclass).

For the semantic roles, two modifier features are always considered important: the average number of modifiers (modmod) and the average number of secondary predicates (modprd).

If we compare the features that are selected in the gold-standard setups to the ones selected in the automatic setup, it can be observed that the added value of the semantic features is more outspoken in the golden setup, especially for the binary task. In the binary classification, 14 golden features are always turned on (= 100) versus 11 automatic ones and in the multiclass setup 19 golden versus 18 automatic features. If we zoom in on the differences between the binary gold versus automatic setup, we observe that the interplay between the features is also different, i.e. for no less than 12 features, the importance is flipped when comparing automatic to gold standard (e.g. for the coreference features the chainspan and corefs have an opposite importance). Again, in the multiclass setup this difference is less outpsoken (6 features have an opposite importance).

6.3 Conclusion

The main focus of this part has been on investigating whether text characteristics based on deep semantic processing – coreference resolution and semantic role labeling – help for automatic readability prediction using supervised machine learning.

We have explained that in order to investigate this we first needed to build a general readability prediction system. We have described in close detail how a corpus was built consisting of a large variety of text material and how this corpus has been assessed on readability. In this respect, we have shown that crowdsourcing is a viable alternative to using expert labels for assessing readability.

Regarding the actual readability prediction system, we explained which new features in the form of five coreference and twenty semantic roles features were implemented together with other features encoding traditional, lexical, syntactic and other semantic information. As prediction task, we defined two classification tasks: a binary and a multiclass task. We explored the added value of our two semantic layers using both standard experiments where these features were manually in- or excluded and joint optimization experiments using a wrapper-based feature selection system based on genetic algorithms. In both setups, we investigated whether there was a difference in performance when these features were derived from gold standard or automatic information.

Our results revealed that readability classification definitely benefits from incorporating semantic information in the form of coreference and semantic role features. The best results for both tasks were achieved after jointly optimizing using genetic algorithms. Contrary to our expectations, we observed that the upper bound of our system was achieved when relying on the automatically predicted deep semantic features. This is an interesting result, because in the end we want to be able to predict readability based exclusively on automatically-derived information sources. In the work performed by Feng et al. (2010), the added value of features derived from coreference information could not be corroborated due to the high level of errors produced by the system for creating these. Though we are fully aware that our results are only valid for the experiments and datasets presented in this dissertation, we can state that for our study the features derived from automatically predicted coreference resolution do seem to contribute to the upper bound of our system.

Part II

Aspect-Based Sentiment Analysis

CHAPTER 7

Preliminaries

In this part, we shift our focus to the domain of sentiment analysis, also known as opinion mining. This is a relatively new strand of NLP research, concerned with modeling subjective information in text. The field has seen rapid expansion in recent years, not only for its many potential end-user applications, but also because of the scientific challenges involved in solving the task. As the level of analysis has become more fine-grained, so has the need for better ways to model subjective text. We explore the contribution of deep semantic processing in the form of automatic coreference resolution and semantic role labeling for fine-grained sentiment analysis.

7.1 Introduction

In today's information society, it cannot be ignored that large parts of our lives are spent and shared online. Originally, the internet was used for communication between a limited number of early adopters only, but with the arrival of Web 2.0 techniques, online communication has become commonplace. Before, the web consisted mainly of static websites and the number of online content consumers far outnumbered the number of content producers. At that time, the web was used mainly to publish and look for (factual) information.

All this changed with the arrival of Web 2.0 sites, which allow site visitors to add content, called *user-generated content* (Moens et al. 2014), thus blurring the boundary between providers and consumers. Examples include forums and message boards, blogs, review sites, e-commerce platforms, but also social networking sites such as Facebook or Twitter. Not only are these a new means of interpersonal and community-level communication, they have also become an important resource for gathering subjective information.

When we need to make a decision about the purchase of a car or cell phone, a travel destination to go to, or a good restaurant to visit, we are typically interested in what other people think. Before Web 2.0, we asked for opinions from friends and family. With the explosive growth of the user-generated content on the Web in the past few years, however, it has become possible to go online and find recommendations or check the experience of other customers, e.g. for a particular restaurant to have lunch at. Instead of relying on anecdotal evidence from friends, we have access to a handy overview of the main aspects of that restaurant enabling us to answer that one crucial question: 'Will I like it?'

The same applies from the perspective of companies, governments and organizations. To know the sentiments of the general public towards its brand, products, policies etc., an organization no longer needs to resort to opinion polls or surveys. Most of that information is already available online, in the form of user-generated content. In previous studies, user-generated content has been used by companies to track how their brand is perceived by consumers (Zabin and Jefferies 2008), for market prediction (Sprenger et al. 2014) or to determine the sentiment of financial bloggers towards companies and their stocks (O'Hare et al. 2009); by individuals who need advice on purchasing the right product or service (Dabrowski et al. 2010) and by nonprofit organizations, e.g. for the detection of suicidal messages (Desmet 2014).

One of the main problems with all data that is produced online is that it is mostly unstructured. Unstructured data is not organized into pre-defined fields (as is the case in a database or table) or annotated for structure (as can be the case in XML). As a consequence, it is not easily interpretable by machines, cannot be readily summarized, etc. The most common example of unstructured data is free text, such as the natural language produced by online users. The main effort to represent the data in this free text in a structured manner is known as the field of the Semantic Web. It is a set of standards that should allow online data to be described consistently, linked to other resources (such as ontologies or Linked Open Data repositories), and exchanged. However, according to its initiators, the effort 'remains largely unrealized' (Shadbolt et al. 2006). In 2010, 90% of all data in the digital universe was unstructured and in 2015, 68% of all unstructured data will be created by consumers (Gantz and Reinsel 2011). This is why natural language processing is still essential for the automatic interpretation of free text.

Typical for user-generated content is that it contains a lot of subjective material. As the amount of online information has grown exponentially, so has the interest in new text mining techniques to handle and analyze this growing amount of subjective text. One of the main research topics is sentiment analysis, also known as opinion mining. The objective of sentiment analysis is the extraction of subjective information from text, rather than factual information. Originally, it focused on the task of automatically classifying an entire document as positive, negative or neutral (i.e. when both types of sentiment are present, or none of them). Typical examples of a document would be a web page, review, comment etc. Liu (2012) summarizes the objective of the field as to analyze “people’s opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes”.

More recently, the focus in sentiment analysis has shifted from coarse-grained opinion mining to fine-grained sentiment analysis, where the sentiment is assigned at the clause level (Wilson et al. 2009). Often, users are not only interested in people’s general sentiments about a certain product, but also in their opinions about specific features, i.e. parts or attributes of that product. Aspect-based (or feature-based) sentiment analysis (Pontiki et al. 2014) focuses on the detection of all sentiment expressions within a given document and the concepts and aspects (or features) to which they refer. Such systems do not only try to distinguish the positive from the negative utterances, but also strive to detect the target of the opinion, which comes down to a very fine-grained sentiment analysis task.

In this part we focus on the task of aspect-based sentiment analysis. We explore whether the availability of more semantic information on discourse entities and their roles helps to pinpoint the words invoking different aspects and the classification of these aspect terms into aspect categories. In order to investigate this in close detail, we first of all required a corpus comprising opinionated text on various aspects of a certain experience or product. Since such a corpus did not yet exist for Dutch, we compiled and annotated a corpus of restaurant reviews, which is presented in Chapter 8. In Chapter 9, we explain how we built the first aspect-based sentiment analysis system able to handle Dutch restaurant reviews and which experiments were conducted in order to both incorporate and assess the added value of our two deep semantic layers: coreference and semantic roles. The results are presented in Chapter 10. For English, a first prototype of our aspect-based sentiment analyser has been described and published in De Clercq et al. (2015). We start by defining sentiment analysis and presenting related research, with a specific focus on aspect-based sentiment analysis and studies where semantic information has been incorporated into the pipeline.

7.2 Definition

Several surveys of the field of sentiment analysis are available, such as the one by Shanahan et al. (2006) or Pang and Lee (2008). However, the book by Liu (2012) is a more recent and extensive summary of this rapidly evolving field.[1] This seminal work offers a comprehensive definition of what an *opinion* is:

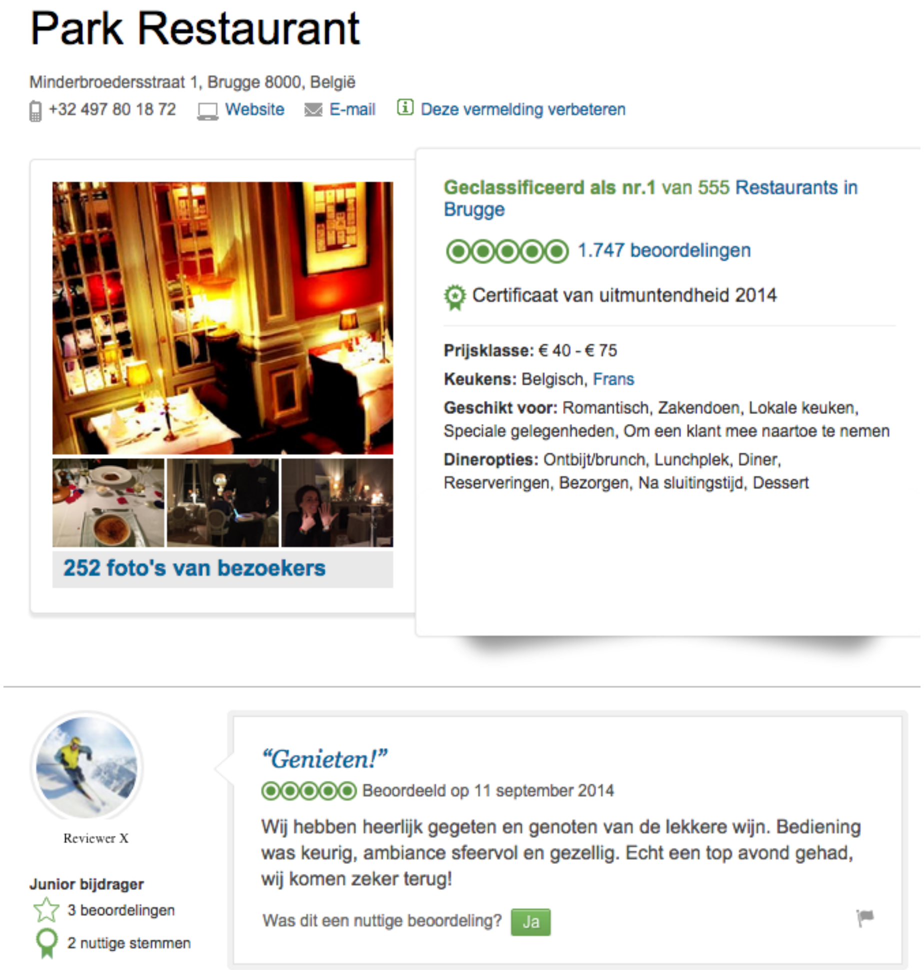
> An opinion is a quintuple, $(e_i, a_{ij}, s_{ijkl}, h_k, t_l)$, where e_i is the name of an entity, a_{ij} is an aspect of e_i, s_{ijkl} is the sentiment on aspect a_{ij} of entity e_i, h_k is the opinion holder, and t_l is the time when the opinion is expressed by h_k. The sentiment s_{ijkl} is positive, negative, or neutral, or expressed with different strength/intensity levels (Liu 2012, 19-20).

Following the definition of Liu (2012), sentiment analysis thus consists of automatically deriving these opinion quintuples from texts and it comprises various subtasks. We will now explain each of these tasks based on an example review presented in Figure 7.1[2].

1. **Entity extraction and categorization**: Extract all entity expressions in a document collection, and categorize or group synonymous entity expressions into entity clusters (or categories). Each entity expression cluster should indicate a unique entity e_i. In our example, the collection consists of restaurant reviews and the entity presented here is ‘Park Restaurant’, belonging to the category *Restaurants*.
2. **Aspect extraction and categorization**: Extract all aspect expressions of the entities, and categorize these aspect expressions into clusters. Each aspect expression cluster of entity e_i represents a unique aspect a_{ij}. These aspects can be both explicit and implicit. In our example, we can find out which aspects of this restaurant are mentioned while reading through the review. The explicit aspects are ‘wijn’ (wine), ‘bediening’ (service) and ‘ambiance’ (ambience). Implicitly, the verbform ‘gegeten’ (literally, have

[1] A query for all papers with ‘sentiment analysis’ in the title, reveals no less than 1700 hits for the range 2010-2015 in Google Scholar and 260 hits in the Web of Science (both were accessed on 10 February 2015)

[2] This example is derived from the corpus of Flemish restaurant reviews that was collected for this dissertation, cfr. Section 8.1.



Park Restaurant

Minderbroedersstraat 1, Brugge 8000, België

+32 497 80 18 72 Website E-mail Deze vermelding verbeteren

Geclassificeerd als nr.1 van 555 Restaurants in Brugge

●●●●● 1.747 beoordelingen

Certificaat van uitmuntendheid 2014

Prijsklasse: € 40 - € 75

Keukens: Belgisch, Frans

Geschikt voor: Romantisch, Zakendoen, Lokale keuken, Speciale gelegenheden, Om een klant mee naartoe te nemen

Dineropties: Ontbijt/brunch, Lunchplek, Diner, Reserveringen, Bezorgen, Na sluitingstijd, Dessert

252 foto's van bezoekers

Reviewer X

Junior bijdrager

☆ 3 beoordelingen

2 nuttige stemmen

"Genieten!"

●●●●● Beoordeeld op 11 september 2014

Wij hebben heerlijk gegeten en genoten van de lekkere wijn. Bediening was keurig, ambiance sfeervol en gezellig. Echt een top avond gehad, wij komen zeker terug!

Was dit een nuttige beoordeling? Ja

EN: We had a wonderful meal and enjoyed the tasty wine. The service was fine, the ambience pleasant and cosy. We truly had a wonderful evening and will definitely return.

Figure 7.1: Review from a particular restaurant in Flanders that was posted on TripAdvisor.

eaten) implies the aspect *Food*. The final sentence is even more implicit and says something about the pleasant experience of going to this restaurant in general. If we classify all these aspect expressions into categories, these could be: *Food*, *Drinks*, *Service*, *Ambience* and *Restaurant* respectively.

3. **Opinion holder extraction and categorization**: Extract opinion holders – h_k – for opinions from text or structured data and categorize them. In our example this can easily be derived from the metadata accompanying the review, i.e. we know who wrote the review. Because of privacy concerns the username was anonymized to 'Reviewer X'.
4. **Time extraction and standardization**: Extract the times when opinions are given and standardize different time formats, t_l. This information can also be easily derived from the time stamp attached to the review: the review was written on September 14, 2014.
5. **Aspect sentiment classification**: Determine whether an opinion on an aspect a_{ij} is positive, negative or neutral, or assign a numeric sentiment rating to the aspect, s_{ijkl}. We can derive that the meal and wine were evaluated positive, as well as the service and ambience, and in the last implicit sentence a positive feeling is expressed in two ways: the evening was great and the opinion holder claims that he will definitely return.

The generated quintuples from our example are:

- (Park Restaurant, *Food*, positive, Reviewer X, Sept-11-2014)
- (Park Restaurant, *Drinks*, positive, Reviewer X, Sept-11-2014)
- (Park Restaurant, *Service*, positive, Reviewer X, Sept-11-2014)
- (Park Restaurant, *Ambience*, positive, Reviewer X, Sept-11-2014)
- (Park Restaurant, *Restaurant*, positive, Reviewer X, Sept-11-2014)

This framework is often called aspect-based or feature-based sentiment analysis (Hu and Liu 2004, Liu et al. 2005, Liu 2012) and is also at the heart of the present investigation.

For the research presented here, the focus is on restaurants reviews which are similar in form to the one presented in our example. Since we can automatically derive the entity, opinion holder and time from the metadata, our main focus will be on the second and fifth subtasks. Actually, we believe this second task consists of two subsequent tasks: aspect term extraction and aspect term classification. In this respect, we follow the task decomposition as suggested by the organizers of two Semantic Evaluation tasks on aspect-based sentiment analysis: SemEval 2014 Task 4 (Pontiki et al. 2014) and SemEval 2015 task 12 (Pontiki et al. 2015).

7.3 Related research

Existing sentiment analysis systems can be divided into lexicon-based and machine learning approaches. Lexicon-based methods (see Taboada et al. (2011) for an overview) determine the semantic orientation of a piece of text based on the words occurring in that text. Crucial in this respect, are sentiment or subjectivity lexicons allowing to define the semantic orientation of words. Lexicons comprise various sentiment or opinion words together with their strength and overall polarity. The word *wonderful*, for example, indicates a positive sentiment, whereas the word *terrible* has a negative connotation. Many subjectivity lexicons were constructed in the past, mainly for English, such as the well-known MPQA lexicon (Wilson et al. 2005) or SentiWordNet (Baccianella et al. 2010). For Dutch, two subjectivity lists were made publicly available, the pattern (De Smedt and Daelemans 2012) and Duoman (Jijkoun and Hofmann 2009) lexicons.

Machine learning approaches to sentiment analysis make use of classification algorithms, such as Naïve Bayes or Support Vector Machines trained on a labeled dataset (Pang and Lee 2008). This dataset can be extracted from existing resources such as reviews labeled with star ratings (Pang et al. 2002) or manual annotation (Wiebe et al. 2005). Crucial in this respect is the engineering of a set of effective features (Liu 2012). Current state-of-the-art approaches model a variety of contextual, lexical and syntactic features (Caro and Grella 2013), allowing them to capture context and the relations between the individual words.

Both lexicon-based and machine learning approaches have been applied for sentiment analysis on various levels, e.g. at the document level (Pang et al. 2002), paragraph level (O'Hare et al. 2009), sentence level (Li et al. 2010), phrase level (Wilson et al. 2009) and word level (Hatzivassiloglou and McKeown 1997). For each of these levels, coarse-grained as well as fine-grained sentiment analysis can be performed. The latter means that sentiment is not only detected and classified, but that it is analyzed more thoroughly, for example by identifying the source and target, i.e. the topic of the expressed sentiment (Kim and Hovy 2006). Together with the advances in the field came a more detailed task definition, as was presented in Section 7.2.

A substantial amount of research has been dedicated to target detection for feature-based or aspect-based opinion mining in product reviews. Often, users are not only interested in people's general sentiments about a certain product, but also in their opinions about specific features, i.e. parts or attributes of that product. Aspect-based (or feature-based) sentiment analysis (Pontiki et al. 2014) focuses on the detection of all sentiment expressions within a given document and the concepts and aspects (or features) to which they refer. Such systems not only try to distinguish positive from negative utterances, but also strive to detect the target of the opinion. Identifying all entities and aspects in a corpus is a non-resolved problem. The most recent advances in this field were made in the framework of two successive SemEval shared tasks devoted to this subject (Pontiki et al. 2014, 2015). Following the SemEval task description, aspect-based sentiment analysis can be decomposed into three subtasks: aspect term extraction, aspect term aggregation or classification and aspect term polarity estimation. We will now discuss the most influential methods and techniques that were used to tackle these three individual tasks.

For the task of **aspect term extraction** (ATE), the most popular and successful approaches are based on frequency and supervised learning (Liu 2012, Pontiki et al. 2014). Hu and Liu (2004) introduced the task of aspect-based sentiment analysis and constructed the first strong baseline for aspect term extraction by identifying all nouns and noun phrases based on part-of-speech tags and counting frequencies. They only kept the frequent nouns and noun phrases using a frequency threshold. In subsequent research, this method was improved by incorporating pruning mechanisms based on pointwise mutual information, meronymy discriminators (e.g. for the camera class these would be 'camera has', 'camera comes with',...) and exploiting the WordNet hierarchy (Popescu and Etzioni 2005). Another improvement was to only include those noun phrases that occur in sentiment-bearing sentences or in certain syntactic patterns (Blair-Goldensohn et al. 2008) or to use the C-value measure which allows to also extract multi-word aspects (Zhu et al. 2009). Most recently, a combination of this frequency baseline with continuous vector space representations of words (Mikolov et al. 2013) has also proven effective in the work of Pavlopoulos and Androutsopoulos (2014).

Using supervised learning, the most dominant method is to approach the ATE task as a sequential labeling task (Liu 2012). Following the IOB2 notation for Named Entity Recognition (Tjong Kim Sang 2002) the aspect term in the annotated training data is labeled with 'B' indicating the beginning of an aspect term, 'I' indicating the inside of an aspect term and 'O' indicating the outside of an aspect term. The two systems achieving the best performance for this subtask in SemEval 2015 Task 12 used this approach. In Toh and Su (2015) (which was actually based on preliminary work (Toh and Wang 2014)), a classifier was trained using Conditional Random Fields (CRF), and in San Vicente et al. (2015) a designated Named Entity Recognition system was used. Both systems implemented typical named entity features such as word bigrams, trigrams, token shape, capitalization, name lists, etc.

The next task is to group aspect terms into categories, known as **aspect term aggregation**. The majority of existing research was performed grouping similar aspect terms into aspect groups without starting out from a predefined set of aspect categories. The most common approaches are to aggregate synonyms or near-synonyms using WordNet (Liu and Lin 2005), statistics from corpora (Chen et al. 2006, Lin and Wu 2009), or semi-supervised learning, or to cluster aspect terms using (latent) topic models (Titov and McDonald 2008, Brody and Elhadad 2010). In other research domain-specific taxonomies have been used to aggregate related terms or hierarchical relations between aspect terms (Kobayashi et al. 2007). More recently, a multi-granular aspect aggregation method was introduced in the work of Pavlopoulos (2014) by first calculating the semantic relatedness between two frequent aspect terms and then performing hierarchical agglomerative clustering to create an aspect term hierarchy.

All these approaches assume that the list of aspect categories is unknown and has to be aggregated from scratch. In this respect, the task definition as proposed in the two aspect-based SemEval tasks differs in that several predefined and domain-specific categories have to be predicted, thus transforming the aggregation task into a multiclass classification task. The two systems achieving the best results on this individual subtask in SemEval 2015 Task 12 both used classification to this purpose, respectively individual binary classifiers trained on each possible category which are afterwards entered in a sigmoidal feedforward network (Toh and Su 2015) and a single Maximum Entropy classifier (Saias 2015). When it comes to the features that were exploited by these systems especially lexical features in the form of bag-of-words such as word unigrams and bigrams (Toh and Su 2015) or word and lemma unigrams (Saias 2015) have proven successful. The best system (Toh and Su 2015) also incorporated lexical-semantic features in the form of clusters learned from a large corpus of reference data, whereas the second-best (Saias 2015) applied filtering heuristics on the classification output and thus solely relied on lexical information for the classification.

The final task is the task of **aspect term polarity classification**. In the context of aspect-based sentiment analysis, the sentiment polarity has to be determined for each mentioned aspect term of a target entity. As explained at the beginning of this section, existing approaches can be divided into two main categories: lexicon-based and machine learning approaches. According to Liu (2012), the key issue is to determine the scope of each sentiment expression within aspect-based sentiment analysis. The main approach is to use parsing to determine the dependency and other relevant information, as done in Jiang et al. (2011) where a dependency parser was used to generate a set of aspect dependent features, or in Boiy and Moens (2009) where each feature is weighted based on the position of the feature relative to the target aspect in the parse tree. With respect to the SemEval tasks it has been shown that general purpose systems used to classify on the sentence level are very effective, which even seems to hold when testing on out-of-domain data (Pontiki et al. 2015).

Semantic roles were investigated for the first time in the work of Kim and Hovy (2006). They used a FrameNet-based semantic role labeler trained exclusively on opinion-bearing frames, which were manually annotated, to determine the holder and topic of opinion expressions. They found that semantic roles add additional information over basic syntactic functions. Another main conclusion of their work was that topic extraction is a much more difficult task than opinion holder extraction. Considering the latter task of opinion holder extraction, Choi et al. (2006) successfully applied a PropBank-based role labeler to this purpose and at the same time extracted opinion expressions using integer linear programming. More recently, Wiegand and Klakow (2010) also studied semantic roles in relation to opinion holder extraction and showed that, in general, the scope immediately encompassing the candidate opinion holder and its nearest predicate together with the scope of the subclause containing the candidate opinion holder provide the best performance.

Ruppenhofer et al. (2008) argue that semantic role techniques are useful but not completely sufficient for holder and topic identification, and that other linguistic phenomena have to be taken into account as well. This was further explored by Johansson and Moschitti (2013), who studied the relation between opinions expressed in text and proposed features derived from the interdependencies between opinion expressions on the syntactic and semantic level. Applying the same techniques to the recognition of attributes, which is closely related to the subtasks of aspect term extraction and classification in aspect-based sentiment analysis, seemed to help as well. We can thus conclude that existing research on applying semantic roles has focused on the investigation of relations and the extraction of relational features allowing to extract opinion holder and topic. In this part, we will incorporate semantic role information in the form of features representing additional semantic information for the subtask of aspect category classification.

Regarding coreference, many survey studies have claimed that the recognition of coreference is crucial for successful (aspect-based) sentiment analysis (Liu 2012, Feldman 2013). Stoyanov and Cardie (2006) were the first to use coreference resolution features to determine which mentions of opinion holders refer to the same entity. In subsequent work (Stoyanov and Cardie 2008), they introduced an approach to opinion topic identification that relies on the identification of topic-coreferent opinions.

Early research in incorporating basic coreference resolution in sentiment analysis was conducted by Nicolov et al. (2008), who investigated how to perform sentiment analysis on parts of the document around topic terms. They demonstrated that using a proximity-based sentiment algorithm can be improved by about 10%, depending on the topic, when using coreference to augment the focus area of the algorithm. The work by Kessler and Nicolov (2009), though its main focus is on finding which sentiment expressions are semantically related, provided some valuable insights in the necessity of coreference as they found that 14% of the targets expressions that had been manually labeled in their corpus were expressed in the form of pronouns. Ding and Liu (2010) introduced the problem of entity and aspect coreference resolution and aimed to determine which mentions of entities and or aspects a certain pronoun refers to, taking a supervised machine learning approach. Their system learns a function to predict whether a pair of nouns is coreferent, building coreference chains based on feature vectors that model a variety of contextual information about the nouns. They also added two opinion-related features, which implies that they used sentiment analysis for the purpose of better coreference resolution rather than the other way around.

To our knowledge, not much qualitative research has been performed investigating whether the availability of coreference information can actually help to improve the aspect term extraction and classification into categories subtasks of aspect-based sentiment analysis. This is a gap that we hope to fill.

CHAPTER 8

Data collection and annotations

Our main objective is to automatically extract domain-specific terms from opinionated text, classify these into broad aspect categories or groups and derive the polarity of the sentiment expressed towards each of these grouped aspects. This can be summarized as the task of aspect-based sentiment analysis. Similar to the readability prediction experiments that were performed in the first part of this dissertation, we tackled this task using a supervised machine learning approach. Our system comprises three incremental subtasks and our main objective was to test the possible added value of incorporating semantic information in the form of coreference and semantic roles. Our hypothesis is that coreference information is especially useful for pinpointing implicit aspect terms and that additional information about the agents' and entities' semantic roles can be helpful for the grouping or classification of these aspect terms into categories.

This is explained in the next chapter, the current chapter presents the Dutch dataset that was collected and annotated in order to serve as the gold standard against which our system could be evaluated. Section 8.1 presents the corpus that was collected for this study which consists of restaurant reviews. Section 8.2 explains the annotation guidelines that were developed to annotate aspects and sentiments and how the annotation process was operationalized using the BRAT annotation tool. Finally, Section 8.3 lists a number of annotation statistics.

8.1 Restaurant review corpus

Aspect-based sentiment analysis has proven important for mining and summarizing opinions from online reviews (Gamon et al. 2005, Titov and McDonald 2008, Pontiki et al. 2014). Systems were developed for a variety of domains, such as movie reviews (Thet et al. 2010), reviews for electronic products, e.g. digital cameras (Hu and Liu 2004) or netbook computers (Brody and Elhadad 2010), and restaurant reviews (Ganu et al. 2009, Brody and Elhadad 2010). Based on this research, several benchmark datasets were made publicly available, such as the product reviews dataset of Hu and Liu (Hu and Liu 2004) or the restaurant reviews dataset of Ganu et al. (2009).

More recently, parts of these two datasets were extracted and re-annotated for two SemEval shared tasks on aspect-based sentiment analysis, SemEval2014 Task 4 (Pontiki et al. 2014) and SemEval 2015 Task 12 (Pontiki et al. 2015). For Dutch, however, no such benchmark datasets exist to our knowledge.

In the framework of this dissertation such a dataset was created, a domain-specific corpus comprising restaurant reviews from TripAdvisor. According to the TripAdvisor website[1], it is the self-declared largest travel website in the world. TripAdvisor is active in over 45 countries around the world and attracts circa 315 million unique visitors a month. On TripAdvisor, reviews on hotels, flights, holiday resorts, restaurants and destinations can be consulted. Reviews can only be submitted by users after creating a personal user profile, but every review is publicly available. An example of a TripAdvisor review was presented in Figure 7.1.

Belgium does not have a separate TripAdvisor website, but in order to find restaurants in the Dutch-speaking part of Belgium, i.e. Flanders, it sufficed to browse to the Dutch landing page and make a query for restaurants in the region of Flanders. We crawled a collection of 400 individual reviews. Four reviews per restaurant were collected, which means we have reviews available for 100 unique restaurants. In order to ensure a broad variety regarding the sentiment expressed within these reviews, we included a comparable amount of top, average and poor restaurants. This responded to selecting restaurants from the first, middle and bottom hits of the TripAdvisor ranking.

Every review was double-checked on language to confirm that it actually contained Dutch text. In a next step the amount of noise was checked since reviews are a form of user-generated content and many state-of-the-art text processing tools, which have all been developed with standard text in mind, show a significant drop in performance when applied to this type of data. This is for

[1] http://www.tripadvisor.nl/pages/about_us.html

example the case when applying parsing (Foster et al. 2011) or named entity recognition (Ritter et al. 2011) to Twitter data. Typical problems that hinder automatic text processing include the use and productivity of abbreviations, deliberate misspellings, phonetic text, colloquial and ungrammatical language use, lack of punctuation and inconsistent capitalization (De Clercq et al. 2013). In our dataset, we found almost no noise; the biggest problem was the lack of proper punctuation which could hinder automatic sentence splitting.

8.2 Annotation

Annotation guidelines had to be developed that would allow to distinguish the different aspects related to a restaurant visit, viz. the opinions expressed towards specific entities (e.g. *pizza margherita, gin tonic*) and/or their attributes (e.g. *quality, price*) and the polarity expressed towards each of these aspects. To this purpose, we relied on the annotation guidelines that were developed in the framework of two SemEval tasks devoted to aspect-based sentiment analysis of English text (Pontiki et al. 2014, 2015). The annotation guidelines presented below are an adapted version of these SemEval guidelines[2].

8.2.1 Annotation guidelines

Every review is annotated on a sentence per sentence basis. To this purpose, all reviews are sentence-split and tokenized using the LeTs preprocessing toolkit (Van de Kauter et al. 2013). Before annotating the actual aspects, these two steps were first manually checked in order to ensure that correct aspect boundaries are assigned.

Important to note is that we only proceed to annotation when at least some level of subjectivity is present within a sentence.

Let us consider the following examples:

(29) NL: Dit is het vuilste, slechtste restaurant ooit bezocht en wij bezoeken er veel per jaar.
EN: This is the dirtiest, worst restaurant we ever visited and we visit a lot of restaurants a year.

[2] http://alt.qcri.org/semeval2015/task12/data/uploads/semeval2015_absa_restaurants_annotationguidelines.pdf

(30) NL: Vanmiddag gaan eten bij Den Cleynen Keyser.
EN: Went to lunch at Den Cleynen Keyser this afternoon.

Example 29 is clearly a subjective sentence where sentiment is expressed, whereas example 30 is just an informative objective sentence. Only example 29 would thus be annotated. The actual annotation consists of two incremental steps: feature expression annotation and opinion expression annotation.

Feature expression annotation

In this step, we have to extract all targets, also known as aspect expressions and categorize these into clusters or categories.

The **target** is the word or words referring to a specific entity or aspect. These are typically:

- Nouns (*NL: pizza, ober, sfeer,... – EN: pizza, waiter, atmosphere,...*)
- Named Entities (*La Cucina, Julia,...*)
- Multi Word Expressions (*NL: witte wijn, biefstuk met friet,... – EN: white wine, steak and fries,...*)

If the same target appears more than once in a sentence, it is only labeled once. In the next example, this applies to the target 'eend'.

(31) NL: [Eend] was wel ok van smaak maar [aardappeltjes] waren niet krokant gebakken en eend was taai.
EN: The [duck] tasted okay but the [potatoes] were not crispy baked and the duck was chewy.

Only explicit aspect expressions are annotated, as a consequence pronouns are not annotated as separate targets, even if they refer to one as in example 32. These pronouns together with other aspects that are referred to implicitly, as illustrated in examples 33, 34 are added as 'NULL' targets to the annotations. In Section 8.2.2 we explain how this is indicated using our designated annotation tool.

(32) NL: *Hij* werd steeds onvriendelijker, toen ik hem antwoorde dat klant toch nog steeds koning was, werd hij alleen maar kwader.
EN: *He* became more and more unfriendly, when I answered that the customer remains king, he only became more upset.

(33) NL: Absolute aanrader, we keren zeker terug
EN: Highly recommended, we will definitely return!

(34) NL: We hebben heerlijk gegeten.
EN: We had a wonderful meal.

Next, aspect or feature **categories** are assigned. We allowed for the same six main categories as present in the SemEval2015 guidelines:

Ambience refers to the atmosphere or the environment of the restaurant's interior or exterior space (example 35). *Drinks* refer to drinks in general or in terms of specific drinks, drinking options, ... (example 36). *Food* focuses on the food in general or in terms of specific dishes, dining options, ... (example 37). *Location* categories deal with the location of the reviewed restaurant in terms of its position, the surroundings, the view (example 38). The *Restaurant* category is used for opinions evaluating the restaurant as a whole and that do not focus on any of the other categories (example 39). *Service* is chosen when the kitchen, counter service, or the promptness and quality of the restaurant's service are reviewed (example 40).

(35) NL: [La Dolce Pizza] is nooit echt proper en de vloer is er zeer plakkerig.
EN: [La Dolce Pizza] is never really clean and the floor is very sticky.

(36) NL: Lauwe [koffie] is gewoon niet accepteerbaar.
EN: Lukewarm [coffee] is just inacceptable.

(37) NL: Ik nam een [tortellini met pancetta] aangeprezen door de uitbater, superlekker!
EN: I took the [tortellini with pancetta], recommended to me by the boss, supergood!

(38) NL: Heel gezellig restaurant in een fantastische [omgeving] in de bossen.
EN: Very cosy restaurant in superb [surroundings] in the woods.

(39) NL: Gewoonweg het beste [restaurant] in de buurt!
EN: Simply the best [restaurant] in the neighborhood!

(40) NL: Maar het [personeel] is er wel vriendelijk.
EN: But the [personnel] is friendly.

Besides these six main categories, various attributes or subcategories also have to be assigned to these main categories, allowing for a more fine-grained annotation. In total, there are five different attributes, which, however, cannot be assigned to all different main categories. Table 8.1 illustrates which attributes can be combined with which main categories.

ASPECT	ATTRIBUTE	EXAMPLES
Ambience	General	NL: Heel gezellig [restaurant] in een fantastische omgeving in de bossen. EN: Very cosy [restaurant] in superb surroundings in the woods.
Drinks	Prices	NL: [Drankjes] zijn abnormaal duur. EN: [Drinks] are priced abnormally high.
	Style & Options	NL: Er is een uitstekende [wijnkaart] die regelmatig vernieuwd wordt. EN: There's an excellent [wine selection] which is updated regularly.
	Quality	NL: Lekkere [wijn] en hele vriendelijke bediening. EN: Tasty [wine] and very friendly service.
Food	General	NL: Jammer, maar dit [gerecht] is niet geslaagd. EN: Too bad, but this [dish] is just not right.
	Prices	NL: De [focaccia] was lekker, maar dat mag wel voor een fortuin. EN: The [focaccia] was tasty, but it should be since it costs a fortune.
	Style & Options	NL: Voor ons beiden kregen we n erg karige [portie frietjes]. EN: We received one very small [portion of fries] for the both of us.
	Quality	NL: Ik nam een [tortellini met pancetta] aangeprezen door de uitbater, superlekker! EN: I took the [tortellini with pancetta], recommended to me by the boss, supergood!
Location	General	NL: Heel gezellig restaurant in een fantastische [omgeving] in de bossen. EN: Very cosy restaurant in superb [surroundings] in the woods.
Restaurant	General	NL: Gewoonweg het beste [restaurant] in de buurt! EN: Simply the best [restaurant] in the neighborhood!
	Prices	NL: Bij [New Havana] te knokke, kan men lekker eten, prijzen liggen wel wat hoger. EN: At [New Havana] in Knokke the food is good but the prices are a bit higher.
	Miscellaneous	NL: Positieve nood voor de hondenliefhebbers, hondjes mogen mee in [restaurant]. EN: Positive aspect for dog lovers, dogs are allowed inside the [restaurant]
Service	General	NL: Maar het [personeel] is er wel vriendelijk. EN: But the [personnel] is friendly.

Table 8.1: Possible aspect main–attribute combinations illustrated with examples. The targets are indicated in between square brackets.

The *General* attribute label is assigned to sentences that express general positive or negative sentiment about an aspect. *Prices* refer to the prices of the food, drinks or restaurant in general. The attribute *Quality*, refers to the taste, the freshness, the texture, the consistency, the temperature, the preparation, the authenticity, the cooking or general quality of the food or the drinks. Another attribute is *Style & Options*, which can be used to refer to the presentation, serving style, the size of the portions, the different options and variety of the food and drinks. Finally, the *Miscellaneous* attribute can be assigned to anything that does not fit into one of the aforementioned cases.

One possible target can be assigned to various categories, if we consider the following example:

(41) NL: De [focaccia] was lekker, maar dat mag wel voor een fortuin.
EN: The [focaccia] was tasty, but it should be since it costs a fortune.

the target [focaccia] should be assigned to two aspect main–attribute categories, namely *Food–Quality* and *Food–Prices*.

Opinion expression annotation

In the next step, we assigned the polarity of the sentiment expressed towards every annotated feature expression. Three main polarities can be distinguished: *positive*, *neutral* and *negative*. The neutral label applies to mildly positive or negative sentiment (example 42) or when two opposing sentiments towards one feature expression occur within one sentence (example 43). The two opposing polarities can be subdivided into a basic *negative*, *positive* (examples 44, 45) or a more intense *very_negative* and *very_positive* version (examples 46, 47).

(42) NL: Viel redelijk mee, maar aan [serranoham met stokbrood] kan niet veel tegenvallen.
EN: Was okay, but then again not much can turn out wrong about [baguette with Spanish serrano ham].
Food–Quality

(43) NL: De [service] is onpersoonlijk maar wel snel.
EN: The [service] did not have a personal touch, but it was fast.
Service–General

(44) NL: De [tomatensoep] was lekker.
EN: The [tomato soup] was tasty.
Food–Quality

(45) NL: [Personeel] onvriendelijk, zelf 2 keer gevraagd om kaart.
EN: [Personnel] unfriendly, had to ask for the menu two times.
Service–General

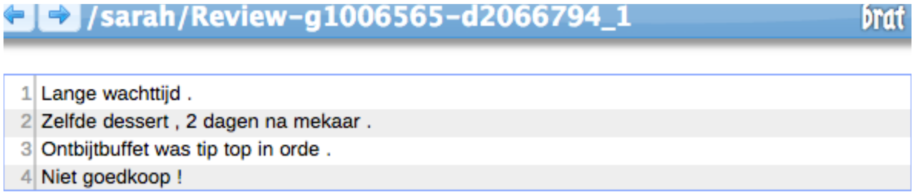
(46) NL: Fantastisch lekker [eten]!
EN: Absolutely divine food!
Food–Quality

(47) NL: [Bediening] zeer slecht, terwijl er niet veel klandizie was.
EN: The [service] is very bad, though there weren't many customers.
Service–General

8.2.2 Annotation tool

All annotations were performed using BRAT[3], the brat rapid annotation tool (Stenetorp et al. 2012). It takes UTF8-encoded text files as input, and stores the annotations in a proprietary standoff format.

Figures 8.1 to 8.5 exemplify how the annotation process is operationalized using BRAT. An annotator loads in a review, as illustrated in Figure 8.1, the review is sentence-split and tokenized. Following the guidelines, the annotator only indicates feature and opinion expressions when sentiment is explicitly being expressed.



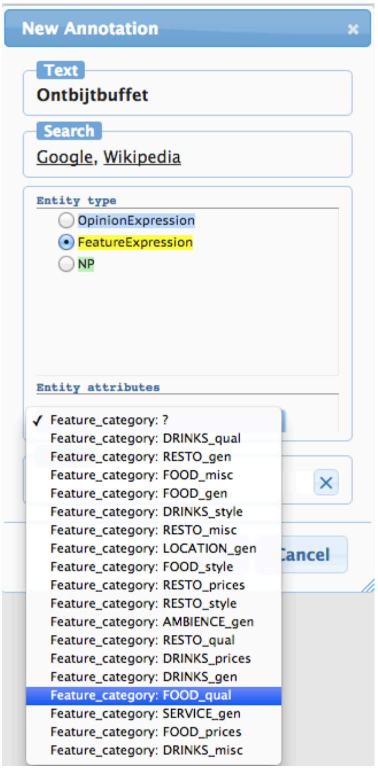
1 Lange wachttijd .
2 Zelfde dessert , 2 dagen na mekaar .
3 Ontbijtbuffet was tip top in orde .
4 Niet goedkoop !

Figure 8.1: Main view of the BRAT interface containing an example review. EN: Long waiting time. Same dessert, two days in a row. Breakfast buffet was spic and span. Not cheap!

When the annotator wishes to annotate a feature expression, first the target is labeled by selecting the appropriate words and then one of the possible aspect categories is assigned (Figure 8.2).

When a target is referred to implicitly this can be indicated by selecting all words in the sentence, assigning one of the possible aspect categories and then typing in the word 'NULL' in the notes section at the bottom of the annotation box, as illustrated in Figure 8.3.

[3] Available at http://brat.nlplab.org/



New Annotation ×

Text
Ontbijtbuffet

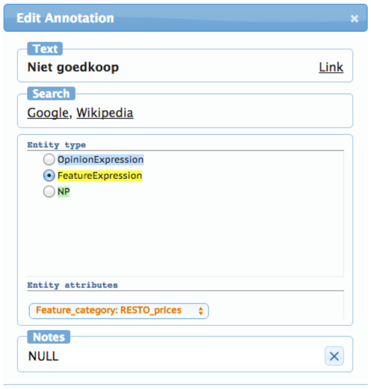
Search
Google, Wikipedia

Entity type
○ OpinionExpression
● FeatureExpression
○ NP

Entity attributes

- ✓ Feature_category: ?
- Feature_category: DRINKS_qual
- Feature_category: RESTO_gen
- Feature_category: FOOD_misc
- Feature_category: FOOD_gen
- Feature_category: DRINKS_style
- Feature_category: RESTO_misc
- Feature_category: LOCATION_gen
- Feature_category: FOOD_style
- Feature_category: RESTO_prices
- Feature_category: RESTO_style
- Feature_category: AMBIENCE_gen
- Feature_category: RESTO_qual
- Feature_category: DRINKS_prices
- Feature_category: DRINKS_gen
- Feature_category: FOOD_qual
- Feature_category: SERVICE_gen
- Feature_category: FOOD_prices
- Feature_category: DRINKS_misc

Figure 8.2: How to indicate a feature expression.



Edit Annotation ×

Text
Niet goedkoop Link

Search
Google, Wikipedia

Entity type
○ OpinionExpression
● FeatureExpression
○ NP

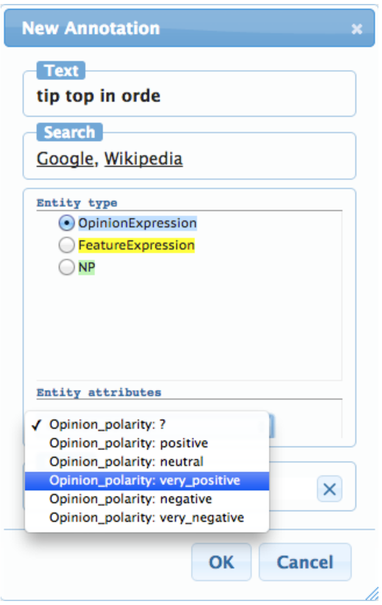
Entity attributes
Feature_category: RESTO_prices

Notes
NULL ×

Figure 8.3: How to indicate an implicit feature expression.

Next, the annotator can label opinion expressions by selecting the words carrying subjectivity and again choosing one of the possible polarity labels (Figure 8.4).

The link between a feature and opinion expression is added by drawing an 'is_about' relation, as exemplified by the orange 'is_about' arrows that are drawn between the feature and opinion expressions in Figure 8.5. Another possible relation is the 'in_span_with' relation which can be drawn between two opinion expressions to indicate that they deal with the same feature expression as exemplified in the second and final sentence of our annotated example.



New Annotation

Text: tip top in orde

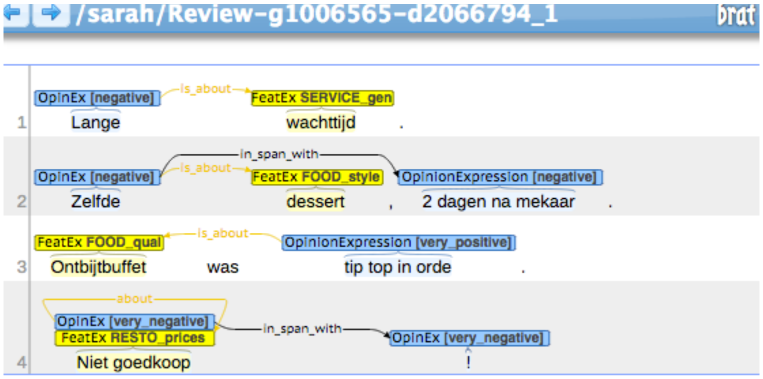
Search: Google, Wikipedia

Entity type:
- ● OpinionExpression
- ○ FeatureExpression
- ○ NP

Entity attributes

✓ Opinion_polarity: ?
Opinion_polarity: positive
Opinion_polarity: neutral
Opinion_polarity: very_positive
Opinion_polarity: negative
Opinion_polarity: very_negative
OK
Cancel

Figure 8.4: How to indicate an opinion expression.



/sarah/Review-g1006565-d2066794_1

1 Lange wachttijd .

2 Zelfde dessert , 2 dagen na mekaar .

3 Ontbijtbuffet was tip top in orde .

4 Niet goedkoop !

Figure 8.5: Main view containing a fully-annotated example review.

8.3 Annotation statistics

Table 8.2 gives an overview of the main corpus statistics: the 400 reviews consist of 2297 sentences and 32564 tokens. On average, every review contains around six sentences and every sentence around fourteen tokens.

	Total	**Average**	**Min**	**Max**
Sentences	2297	5.74	1	41
Tokens	32546	14.17	2	112

Table 8.2: Data statistics of our restaurants review corpus (total number of sentences and tokens in all reviews and average, minimum and maximum value of respectively sentences per review and tokens per sentence).

All 2297 sentences were manually annotated with feature and opinion expressions by a trained linguist. These annotations were verified by another linguist and disagreement was resolved through mutual consent.

If we consider the following example:

(48) NL: We hebben allebei [pizza frutti di mare] gegeten en deze was qua prijs/kwaliteit best ok.
EN: We both took the [pizza frutti di mare] and it was okay with regard to the price/quality ratio.

there was disagreement whether the target [pizza frutti di mare] should receive the aspect category *Food–Quality*, *Food–Prices*, both or maybe the label *Food–General*. We decided that the aspect category *Food–General* was the best option since the review expresses something about the food in general and does not explicitly focus on either the quality or price. Inter-annotator agreement (IAA) on this particular dataset was not calculated but for the English guidelines on which these guidelines were based, a Dice coefficient of 0.72 for the target labeling and a Kappa value of 75.34 for the polarity classification were reported in Pavlopoulos (2014).

Out of the 2297 sentences, 76% (n = 1767) were considered as subjective, whereas 34% (n = 530) as not opinionated at all. The opinionated sentences were further annotated with feature and opinion expressions. In total, 2445 targets were annotated, ranging from sentences including one to twelve individual targets. Implicit targets or pronouns referring to aspect categories were also labelled by adding them as 'NULL' target to the annotations (as briefly discussed in Section 8.2.1). In total, 31.6% (n= 773) of the annotated targets were implicit.

In Table 8.3, we give an overview of how many of each of the aspect or feature categories are present in our data – based on both explicit and implicit ('NULL') targets – together with the amount of positive (POS), neutral (NEUT) or negative (NEG) sentiment expressions. To this purpose we merged the more intense polarity classes (very_positive and very_negative) with their global counterparts (positive and negative).

FEATURE EXPRESSIONS			OPINION EXPRESSIONS		
Main	**Attribute**	**#**	**POS**	**NEUT**	**NEG**
Ambience	General	240	169	18	53
Drinks	Prices	23	7	2	14
	Style & Options	38	15	3	20
	Quality	68	54	1	13
Food	General	15	7	5	3
	Prices	54	24	5	25
	Style & Options	209	103	16	90
	Quality	675	448	60	167
Location	General	34	24	4	6
Restaurant	General	437	274	21	142
	Prices	43	16	5	22
	Miscellaneous	26	7	5	14
Service	General	583	260	37	286

Table 8.3: Annotation statistics representing the amount of annotated feature and opinion expressions in our restaurant review dataset.

If we consider our six main aspect categories, we notice that three main aspect categories – *Food*, *Restaurant* and *Service* – are mentioned most often, as visualized in Figure 8.6.

If we have a closer look at these three main aspect categories also allowing for a more fine-grained attribute labeling, we notice that when it comes to the *Food* or *Drinks* aspects, especially the *Quality* seems important to mention. For the *Restaurant* itself, however, more focus is given to the *General* feeling people had when dining in a particular restaurant and whether they will return or not.

When we investigate the distribution of the opinions expressed towards each of the main features, as visualized in Figure 8.7, we clearly notice that in our dataset there are overall more positive opinion expressions, visualized by the green bars. Especially when people refer to more general aspects such as the *Ambience* (70% positive polarity) in a restaurant or the *Location* (71% positive

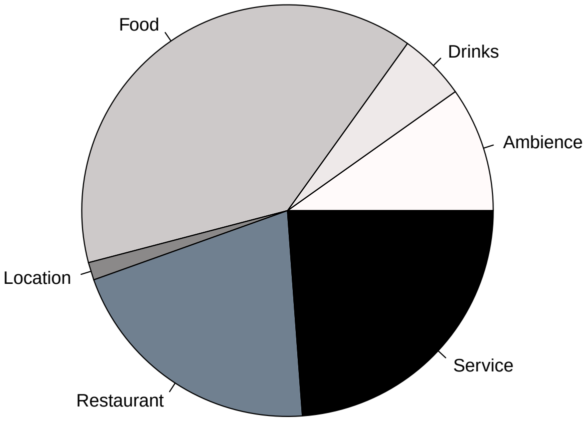


Figure 8.6: Pie chart visualizing the main category distribution in our dataset.

polarity), people tend to make positive remarks in our dataset. Only for the aspect category *Service*, we observe that when people comment about the service, they express slightly more negative (49%) than positive (45%) feelings.

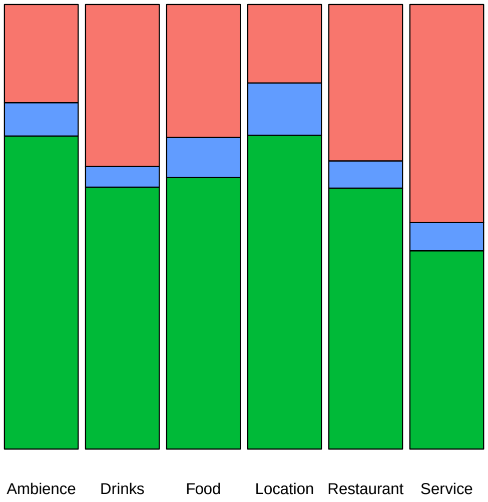


Figure 8.7: Heat barplots visualizing the amount of positive (green), neutral (blue) and negative (red) opinions expressed within each main aspect category.

To conclude, we discuss the 'NULL' targets in isolation which account for more than 30% of the annotated targets. We hypothesize that these will present a challenge to any automatic aspect-based sentiment analysis system.

Considering the classification into main–attribute categories (Fig. 8.8), more than half of the implicit target mentions refer to the main *Restaurant* category, i.e. 341 out of 773 targets, followed by *Service* (186) and *Food* (173). If we look at the more fine-grained attributes of the largest main category, i.e. *Restaurant*, especially the *General* attribute is referred to implicitly, i.e. 296 times. When we compare this to the overall amount of mentions expressed towards this main–attribute category (see Table 8.3), we can conclude that more than half (296 out of 415 *Restaurant–General* aspect expressions) are implicit.

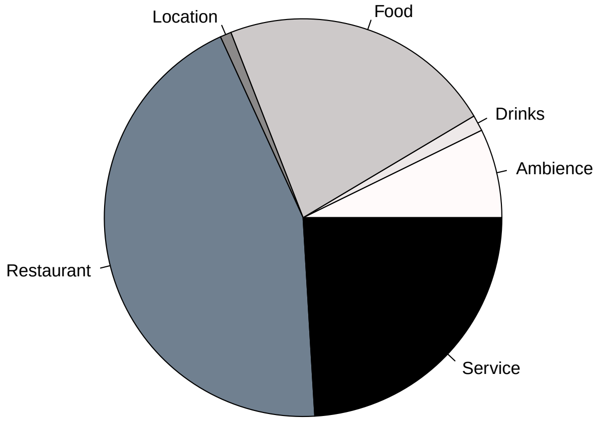


Figure 8.8: Pie chart visualizing the main category distribution of the implicit targets in our dataset.

Considering the opinions (Fig. 8.9), in the most popular main aspect category, *Restaurant*, the feelings expressed remain overall positive (58%). The negative opinions towards the aspect category *Service*, however, become more outspoken when referred to implicitly (72%).

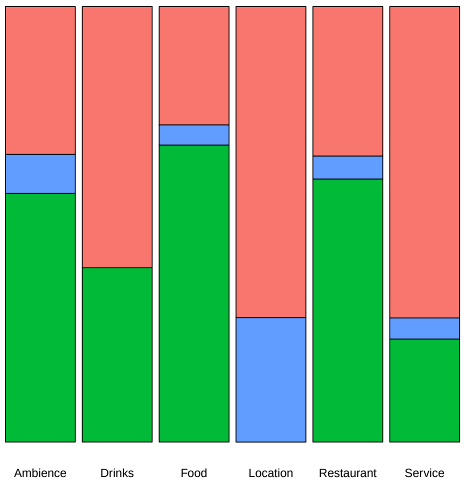


Figure 8.9: Heat barplots visualizing the amount of positive (green), neutral (blue) and negative (red) opinions expressed within each implicit main aspect category.

CHAPTER 9

Experiments

In this chapter, we present our aspect-based sentiment analysis pipeline and explain how we explored the contribution of deep semantic processing in the form of automatic coreference resolution and semantic role labeling. Following the SemEval subtask classification (Pontiki et al. 2014, 2015), we discerned three individual subtasks:

1. **Aspect term extraction**. In a first step, candidate terms have to be automatically selected from running text. An additional constraint is that candidate terms can only be selected if sentiment is expressed towards these terms.
2. **Aspect category classification**. Based on the selected candidate terms, in this multiclass classification task a distinction has to be made between thirteen possible aspect categories (cfr. Table 8.1).
3. **Aspect polarity classification**. The final step consists of a multiclass classification task where the polarity towards each indicated aspect has to be labeled. We allow a distinction between three possible polarity labels: positive, negative and neutral.[1]

[1] The *very_positive*, *positive* and *very_negative*, *negative* labels were merged.

In a fully-automatic setup, these three subtasks are performed incrementally: first aspect terms are automatically derived, next they are assigned to a correct aspect category, and finally their polarity is classified. This chapter details how we tackled each of these individual tasks.

For the first task, we employed a hybrid terminology extraction system to extract candidate aspect terms from running text and at the same time applied a subjectivity heuristic to discern whether sentiment is actually expressed (Section 9.1). For the second and third task, we adopted a supervised machine learning approach. In order to classify the aspect terms into the different categories, a classifier was trained using a rich feature space (Section 9.2). We included lexical features (derived from token unigrams of the sentence in which an aspect term occurs) and semantic features (derived from the explicit aspect terms). For the final task of aspect polarity classification we relied on a previously developed polarity classifier (Van Hee et al. 2014), which was adapted to deal with Dutch text (Section 9.3).

Our main interest was to test the possible added value of incorporating our two semantic information layers in the form of coreference and semantic roles. Our hypothesis is that coreference information is especially useful for pinpointing implicit aspect terms, which form a substantial part of the annotated aspect terms or targets in our review corpus (i.e. more than 30% as shown in Section 8.3). On the other hand, we assume that additional information about the agents' and entities' semantic roles could be helpful for the grouping or classification of the different aspect terms into the thirteen possible aspect categories.

Figure 9.1 visualizes the architecture that was developed in order to perform the task of aspect-based sentiment analysis (ABSA). The two semantic layers are highlighted in blue to illustrate where these were plugged into the pipeline.

Previous research using our approach on similar English data (De Clercq et al. 2015) and preliminary experiments on our Dutch restaurant corpus, revealed that each individual subtask is prone to a number of errors and would benefit from optimization.

To overcome this, the restaurant review corpus was split in two subsets: a development set comprising 300 reviews and a held-out test set comprising 100 reviews. In this way, the development set was used for optimizing the performance on each individual subtask, whereas the held-out test set was used to test the fully-automatic pipeline. Some statistics regarding these two datasets are presented in Table 9.1

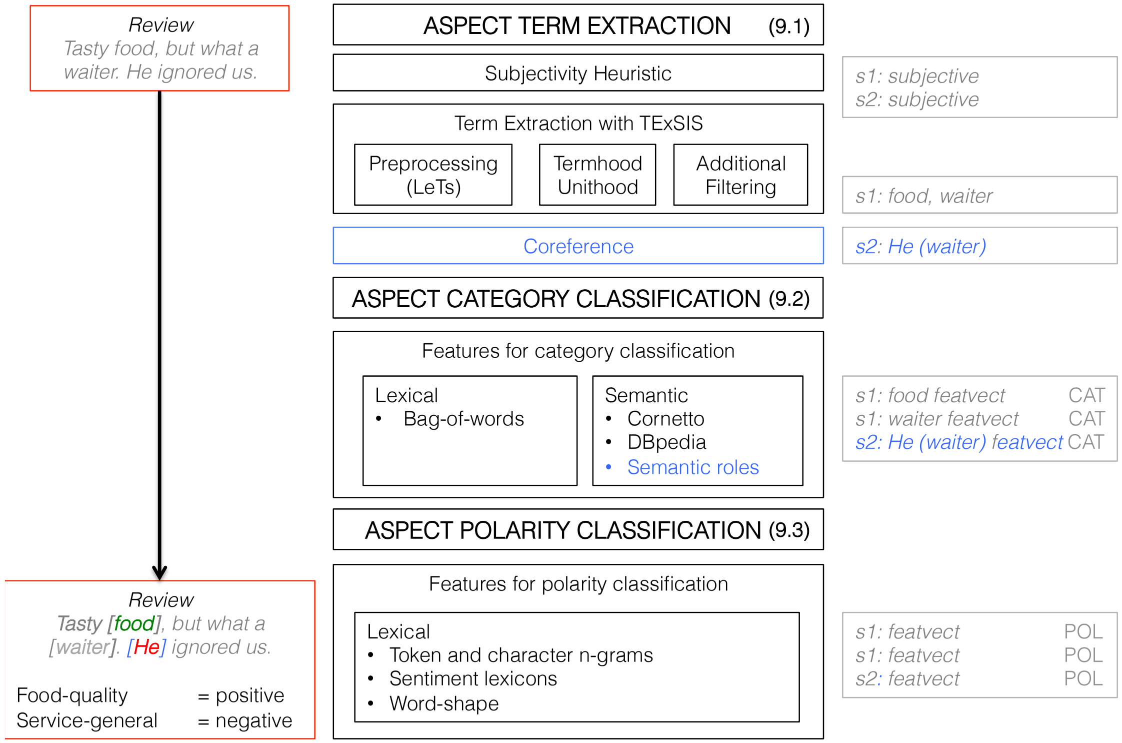


Figure 9.1: Architecture of our ABSA pipeline illustrated with an example. The two semantic layers (coreference and semantic roles) are indicated in blue. The *s* on the right-hand side stands for sentence.

	Development	Held-out
Reviews	300	100
Sentences	1722	575
Tokens	24894	7652
Targets	1843	602
‘NULL’ targets	563	210

Table 9.1: Statistics of our development set and held-out test set. The upper part presents the number of reviews, sentences and tokens each set contains. The lower part indicates how many gold-standard aspect terms or targets are present in total in both datasets and how many of these constitute a ‘NULL’ target.

9.1 Aspect term extraction

Prior to classification, it is essential to know which terms or concepts are present in the review that refer to various aspects accompanying a restaurant visit. An additional constraint, however, is that aspect term expressions could only be extracted when sentiment towards these was clearly expressed. To this purpose, a subjectivity heuristic was implemented as a first step.

9.1.1 Subjectivity heuristic

We filtered the occurrence of subjective words on a sentence-per-sentence basis and implemented this as the first step in our architecture (cfr. Figure 9.1).

Since in this step, especially coverage is of crucial importance, we combined a variety of existing external lexicons and manually created domain-specific lexicons. We thus follow a lexicon-based approach to detect subjectivity (Taboada et al. 2011). Since the same external lexicons were used for the third subtask of polarity classification, these will be explained in closer detail in Section 9.3. Besides these external lexicons, we manually created domain-specific lexicons.

To identify the sentiment words, a lookup was performed for both the surface form and lemma of each word in the sentence. To this purpose shallow linguistic preprocessing was performed on all reviews using the LeTs Preprocess toolkit (Van de Kauter et al. 2013).

9.1.2 Terminology extraction using TExSIS

Wright (1997) defines terms as ‘words that are assigned to concepts used in the special languages that occur in subject-field or domain-related texts’. In order to detect these candidate terms, terminology extraction was applied to determine whether a word (sequence) is a term that characterizes the target domain. In an attempt to define terms, Kageura and Umino (1996) proposed properties such as termhood (the degree to which a linguistic unit is related to domain-specific concepts) and unithood (the degree of strength or stability of syntagmatic combinations and collocations). Unithood is relevant for multi-word terms, whereas termhood deals with both single-word and complex terms. Both linguistic and statistical approaches have been proposed to tackle automatic term extraction (ATE). Linguistic approaches identify terms by their syntactic properties. This is mostly done in a two-step procedure, in which automatic shallow parsing of texts is followed by simple term extraction using e.g. regular expressions. By relying on language-specific information, linguistically-based ATE is language-dependent (see for example Justeson and Katz (1995) for English term-formation patterns). Statistical ATE, on the other hand, is language-independent, using measures such as frequency, association scores, diversity and distance metrics to distinguish true from false terms. We refer to Zhang et al. (2008) for an overview of both methodologies.

For our aspect term extraction, we tested an existing hybrid terminology extraction system, TExSIS (Macken et al. 2013). Whereas a more heuristic approach to aspect term extraction based on the frequency of noun phrases was already introduced by Hu and Liu (2004) (cfr. Section 7.3), it is the first time that an end-to-end terminology extraction system is used for this specific task. TExSIS is a hybrid system in that it combines linguistic and statistical information.

For the linguistic analysis, TExSIS relies on tokenized, Part-of-Speech tagged, lemmatized and chunked data using the LeTs Preprocess toolkit (Van de Kauter et al. 2013). Subsequently, all words and chunks matching certain Part-of-Speech patterns (e.g. Noun, Noun + Noun, Noun + Preposition + Noun, ...) are extracted as candidate terms. In order to determine the specificity of and cohesion between these candidate terms, several statistical filters are combined to determine the termhood and unithood (Kageura and Umino 1996) of a given term. Within TExSIS, the log-likelihood ratio (Rayson and Garside 2000) is calculated on all single-word terms, C-values (Frantzi and Ananiadou 1999) are derived for the multi-words units and in a final step all single and multi-word terms are ranked using the term weighting measure as proposed by Vintar (2010). Since we are dealing with Dutch text, these three statistical filters were calculated using the SoNaR 500 million words corpus (Oostdijk et al. 2013) as a reference corpus.[2] For more details on these statistical filters, we refer to Macken et al. (2013).

TExSIS was developed as a generic terminology-extraction system. We, however, are only interested in very specific terms that are closely related to the restaurant-domain. In previous experiments performed on very similar English data (De Clercq et al. 2015), it was decided to use a reduced version of TExSIS, focussing mainly on the linguistic noun phrase extraction and relying on the term weighting ranking, while, at the same time, applying some domain-specific filtering heuristics.

If we consider the following example:

(49) Na een [goede [aperitief]] bestelde ons [mama] een [[pizza] [margherita]], die was heerlijk!
EN: After a [good [appetizer]] our [mother] ordered a [[pizza] [margherita]], which was divine!

Using the candidate term list outputted by TExSIS as such, our system would normally indicate the six terms between square brackets. Based on our previous experience, however, our system was adapted so that it would always prefer the largest possible unit when multiple candidate terms are possible. For our example, this would leave us with three candidate terms: [goede aperitief], [mama] and [pizza margherita].

The additional domain-specific filtering heuristics that were applied to the list of candidate terms are presented next.

9.1.3 Filtering

We first of all applied **subjectivity filtering** using a subjectivity lexicon. During the Duoman project a lexicon was composed consisting of nouns, adjectives, verbs and adverbs with polarity scores between -1 and 1 (Jijkoun and Hofmann 2009). These scores were determined by bootstrapping from the translation of an English lexicon. In order to evaluate the quality of this automatically created lexicon, a gold standard was created, in which two human annotators marked words as positive, negative or neutral by using a five-point scale ranging from very negative (- -) to very positive (++).

[2] Since SoNaR-500 comprises both standard text material and text coming from new media, this corpus was filtered to only include standard text material (e.g. text coming from newspapers, magazines, etc.)

Since in this step the quality is important, the gold-standard lexicon comprising 2595 terms was used as a filtering lexicon in our system to rule out candidate terms containing subjective words by performing a string match. Considering our running example 49, the candidate term [goede aperitief] would thus be filtered out and replaced with the term [aperitief]. The other two terms [mama] and [pizza margherita] would remain.

In a next step, we applied **semantic filtering**. Semantic annotation deals with enriching texts with pointers to knowledge bases and ontologies (Reeve and Han 2005). This can be done by linking mentions of concepts and instances to either semantic lexicons like WordNet (Fellbaum 1998), or Wikipedia-based knowledge bases (Hovy et al. 2013) such as DBpedia (Lehmann et al. 2013). In aspect-based sentiment analysis the use of WordNet has been investigated before and has proven successful in this respect (Popescu and Etzioni 2005). Data coming from the Linked Open Data-cloud such as DBpedia, however, have not been exploited before. We use both:

- Cornetto (Vossen et al. 2013) is a resource combining and aligning two existing semantic resources for Dutch: the Dutch wordnet (Vossen 1998) and the Referentie Bestand Nederlands (Maks et al. 1999, Martin 2005). The latter is a Dutch database with combinatoric information of Dutch word meanings. Cornetto covers 40K entries. We used this semantic lexicon to perform an additional filtering of not domain-specific candidate terms. To this purpose, we exploited the manual category annotations and, for each main category (*Ambience, Drinks, Food, Location, Service* and *Restaurant*), derived a value indicating the number of (unique) terms annotated as aspect terms from that category that (1) co-occur in the synset of the candidate term or (2) which are a hyponym/hypernym of a candidate term in the synset. In case the candidate term was a multi-word term, the full term of which was not found in the annotations, this value was calculated for all nouns in the multi-word term and the resulting sum was divided by the number of nouns. The values can be perceived as semantic links.

  For our running example, ideally the candidate term [mama] would receive no synonym or hypernym link with one of the main aspect categories, whereas the term [aperitief] and multi-word term [pizza margherita] would receive a semantic link with the main categories *Drinks* and *Service*, respectively.

- DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia and make this information available as linked RDF data (Lehmann et al. 2013). This repository encodes, for example, which objects have which properties, or which relations hold between two given objects. For this research, we automatically identified the concepts and categories of our candidate terms on the basis of a two-step process as visualized below.

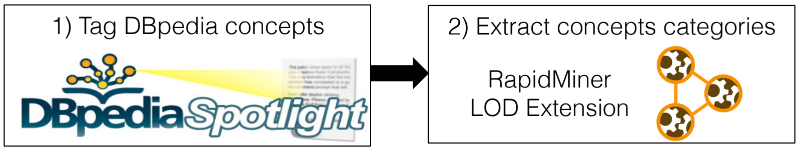


Figure 9.2: Two-step annotation process using semantic web technologies.

First, we identified concepts in DBpedia by processing each candidate term with DBpedia Spotlight (Mendes et al. 2011). Next, categories for each concept were created, corresponding to the categories in Wikipedia. To that end, we extracted all direct categories for each concept (`dcterms:subject`), and added the more general categories with a maximum of two levels up in the hierarchy (`skos:broader`). This process is illustrated in Figure 9.3. The whole process, comprising the annotation with DBpedia Spotlight and the extraction of categories, was performed in the RapidMiner LOD Extension (Paulheim and Fürnkranz 2012).

Applied to our example, we assume the terms [aperitief] and [pizza margherita] might receive a DBpedia category related to the restaurants-domain, whereas the term [mama] not.

9.1.4 Experimental setup

As previously mentioned, two different sets of experiments were conducted for each subtask. In a first set only the development data was used in order to optimize (i.e. 300 reviews) and in a second set the resulting optimal setting was tested on the held-out test data (i.e. 100 reviews).

To evaluate, precision, recall and F-score[3] were calculated, by comparing the list of aspect term expressions or targets that our system returned for a sentence to a corresponding gold-standard list. In this respect, we followed the SemEval2015 evaluation (Pontiki et al. 2015) and used the begin and end position of each target in the sentence to evaluate. This means there has to be an exact match, making this a very strict evaluation. Moreover, it should be noted that this evaluation discards implicit 'NULL' targets since these do not correspond to

[3] See Section 2.3.3 for the formulas.

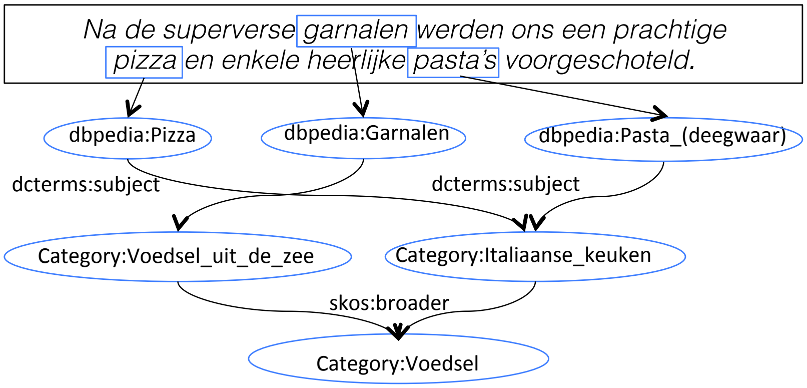

Na de superverse garnalen werden ons een prachtige pizza en enkele heerlijke pasta's voorgeschoteld.
dbpedia:Pizza
dbpedia:Garnalen
dbpedia:Pasta_(deegwaar)
dcterms:subject
dcterms:subject
Category:Voedsel_uit_de_zee
Category:Italiaanse_keuken
skos:broader
Category:Voedsel

Figure 9.3: Example sentence in which the candidate aspect terms are semantically enriched using DBpedia.
EN: After the superfresh shrimps, we were served a wonderful pizza and some delicious pasta dishes!

explicit aspect term expressions. In addition, targets referring to multiple aspect categories within one sentence are only counted once (viz. Example 41 where the term [focaccia] referred to two aspect categories, *Food–Quality* and *Food-Price*).

Previous experiments revealed that using TExSIS for candidate term extraction leads to a reasonable recall but low precision (De Clercq et al. 2015). Through the incorporation of the additional filtering heuristics we thus strived for a better balancing of precision and recall. In order to do this optimization, the development data was split in two subsets – a devtrain set containing 250 reviews and a devtest set containing 50 reviews. This was necessary especially for the semantic filtering, because there matches with the gold-standard annotations were necessary in order to derive this information. In the next round, the resulting optimal setting was then tested on the held-out data comprising 100 reviews. In this latter setup, the semantic filtering was performed on the entire development set.

The focus of this evaluation was on the extraction of explicit aspect terms, ignoring the implicit aspect terms, i.e. the 'NULL' targets which account for more than 30% of the annotated aspect term expressions in our corpus (cfr. Section 8.3). As illustrated in Figure 9.1 and mentioned at the beginning of this chapter, we hypothesized that coreference information is useful for pinpointing implicit aspect terms and can be incorporated immediately after the actual term extraction. As the incorporation of this additional information source could only be evaluated after the aspect terms had been assigned to specific categories, the added value of this step was investigated in the next subtask, i.e. aspect category classification.

9.2 Aspect category classification

Given a list of possible candidate terms, the next step consisted in classifying these terms into broader aspect categories. As discussed in Section 8.2, this refers to both the detection of the six main categories (e.g. *Food*) and the various attributes or subcategories (e.g. *Prices*). See Table 8.1 for an overview of all possible labels.

The two systems achieving the best results on this individual subtask in SemEval 2015 Task 12 both used classification to this purpose, respectively individual binary classifiers trained on each possible category which are afterwards entered in a sigmoidal feedforward network (Toh and Su 2015) and a single Maximum Entropy classifier (Saias 2015). When it comes to the features that were exploited by these systems, especially lexical features in the form of bag-of-words such as word unigrams and bigrams (Toh and Su 2015) or word and lemma unigrams (Saias 2015) have proven important. The best system (Toh and Su 2015) also incorporated lexical-semantic features in the form of clusters learned from a large corpus of reference data, whereas the second-best (Saias 2015) applied filtering heuristics on the classification output and thus solely relied on lexical information for the classification.

Given that this is a fine-grained classification task requiring a system to grasp subtle differences between various main–attribute categories (e.g. *Food–General* versus *Food–Prices* versus *Food–Quality* versus *Food–Style&Options*), we believed that additional semantic information would be crucial besides lexical information. Regarding our two semantic information sources, we hypothesized that information about the agents' and entities' semantic roles could be helpful for this fine-grained classification task and that coreference information is useful for pinpointing implicit aspect terms which, in turn, allows to derive additional semantic information for these implicit terms.

For this classification task, we relied on a rich feature space by extracting lexical bag-of-words features based on the sentence in which a target term occurs and by deriving additional semantic information from the target term itself. Afterwards, multiclass classification into the thirteen domain-specific categories was performed using LibSVM.[4] We were mainly interested in exploring the added value of semantic information in general and our two semantic layers to be more precise.

9.2.1 Lexical features

As a baseline, we derived bag-of-words token unigram features of the sentence in which a term or target occurs in order to represent some of the lexical information present in each of the categories. In bag-of-words representations, each feature corresponds to a single word found in the training corpus.

Bag-of-words features have proven successful in the domains of information retrieval and text classification (Manning et al. 2008). In the SemEval 2015 Task 12 (Pontiki et al. 2015), these features also constituted the baseline for the task of aspect term classification and the two top-performing systems revealed that using this kind of information leads to top performance (cfr. supra).

9.2.2 Lexical-semantic features

An analysis of the top-performing system of the SemEval2015 Task 12, revealed that besides lexical features, features in the form of clusters derived from a large reference corpus of restaurant reviews, are useful (Toh and Su 2015). For Dutch we did not have such a large reference corpus available, but we did include the two semantic information sources that were used for the subtask of automatic term extraction to derive lexical-semantic features.

For Cornetto, this translated to six features (Table 9.2), each representing a value indicating the number of (unique) terms annotated as aspect terms from that category that (1) co-occur in the synset of the candidate term or (2) which are a hyponym/hypernym of a candidate term in the synset.

For DBpedia, the process where the concept and categories of the aspect terms are automatically identified led to a large number of possible categories. Our list of gold-standard targets in the development data, for example, led to 451 unique DBpedia categories. After a manual inspection, eighteen unique categories were included as binary features into our features space, which are listed in Table 9.3. It is clear that each of these categories can be used to generalize to one or more of the six main aspect categories. This means that the distinction of more fine-

[4] Preliminary experiments revealed that performing one-step classification achieved better results than perceiving the subtask as a two-step classification task where first the main category and afterwards the attribute category had to be assigned.

Feature	Explanation
Cor_AMBIENCE	Cornetto match with the main category *Ambience*
Cor_DRINKS	Cornetto match with the main category *Drinks*
Cor_FOOD	Cornetto match with the main category *Food*
Cor_LOCATION	Cornetto match with the main category *Location*
Cor_RESTAURANT	Cornetto match with the main category *Restaurant*
Cor_SERVICE	Cornetto match with the main category *Service*

Table 9.2: Cornetto features corresponding to each of the main aspect categories.

grained main–attribute categories is still left to the bag-of-words representations. This brings us to one of our deep semantic information sources, namely semantic roles.

Feature	Translation
DB_Alcoholische_drank	DB_Alcoholic_beverages
DB_Brood_en_banket	DB_Bread_and_pastry
DB_Broodbeleg	DB_Breadspread
DB_Eetbare_plant	DB_Edible_plant
DB_Gastronomie	DB_Gastronomy
DB_Gerecht	DB_Dish
DB_Horeca	DB_Catering_industry
DB_Keuken_naar_land	DB_Cuisine_per_country
DB_Kruiden_en_specerijen	DB_Herbs_and_spices
DB_Meubilair	DB_Furniture
DB_Niet-alcoholische_drank	DB_Non-alcoholic_beverages
DB_Pluimvee	DB_Poultry
DB_Snoep	DB_Candy
DB_Vee	DB_Cattle
DB_Vis	DB_Fish
DB_Voeding	DB_Nutrition
DB_Voedsel	DB_Food
DB_Voedselterminologie	DB_Food_terminology

Table 9.3: DBpedia features together with their English translation.

9.2.3 Semantic role features

We hypothesized that additional information about agents' and entities' semantic roles could provide additional semantic evidence for resolving the aspect category classification task with regard to the more fine-grained main–attribute labels. In this respect, the predicates evoking certain roles, for example, constitute an added value on top of the bag-of-words features when it comes to discerning the different attributes (e.g. 'The food **tasted** good' versus 'The food just **cost** too much)'.

For the extraction of semantic role features, every review was processed with SSRL trained on the complete 500,000 subset of the SoNaR 1 corpus comprising a variety of text genres (Section 2.2.1). As a result, semantic roles were indicated at the clause level of every sentence within a review. For the feature construction, we retrieved the position of every target term and derived whether the lexical unit comprising this aspect term evokes a semantic role or not.

Consider the following example:

(50) Het dessert en mignardises bij de koffie zorgden voor de perfecte afsluiter.
EN: The dessert and mignardises accompanying the coffee provided the perfect ending.

This sentence contains two aspect expressions, viz. 'dessert' and 'mignardises', each referring to the same aspect category, i.e. *Food–Quality*. Following the SSRL semantic role labeling, the following semantic roles were automatically derived for this sentence:
Het dessert en mignardises bij de koffie –*Arg0*
zorgen –*PRED*
voor de perfecte afsluiter –*Arg1*

This semantic role information is stored in 19 binary features, each representing a possible semantic role label (one predicate, five possible arguments and fourteen possible modifiers[5]), which are added to the feature vector. Finally, in order to investigate whether the actual predicates also carry meaningful semantic information that might allow to better distinguish between aspect categories (e.g. 'The food **tasted** good' or 'The waiter **serving** us was wonderful'), we also included the predicate token as a separate feature.

For our running example 50, this would mean that the two feature vectors built for the instances 'dessert' and 'mignardises' would both contain 18 binary semantic role features with the value zero, and one semantic role feature, i.e. Arg0,

[5] See Section 2.3.1 for a more detailed description of every possible semantic role.

with the value one. In addition, the predicate token 'zorgden' would be added to the feature vector.

9.2.4 Coreference

Following the annotation guidelines (Section 8.2.1), all implicit aspect mentions and pronouns referring to aspects have been annotated as 'NULL' targets. In our corpus more than 30% of the aspects are referred to implicitly (773 out of the 2445 aspect terms). As a consequence, even if we would have gold-standard target mentions at our disposal in the form of 'NULL' targets that are added to our instancebase, the respective values of the lexical-semantic features could not be added because these cannot be derived from 'NULL' targets. Moreover, semantic roles cannot be extracted because no position for these 'NULL' targets can be derived in the sentence.

If we have coreference information available, however, we hypothesize that for certain 'NULL' targets these features can actually be derived. In other words, a coreferential relation between an anaphor – pronoun – and an antecedent constituting an aspect term in itself should enable us to derive additional semantic information.

Let us consider the following example:

(51) In La Dolce Vita bestelden we een lekkere [pizza], ons aanbevolen door de baas. *Die* smaakte gewoonweg fantastisch!
EN: In La Dolce Vita we ordered a tasty [pizza], recommended by the boss. *It* tasted absolutely divine!

In the first and second sentence, an opinion is expressed on an aspect category, i.e. *Food–Quality*, in an explicit and implicit manner, respectively. If coreference information is added to this example, the pronoun 'it' would be found to refer back to 'pizza', thus constituting an anaphor–antecedent pair. This antecedent can then be used to derive additional lexical-semantic information in the form of Cornetto (*Cor_FOOD: 1.0*) and DBpedia features (*DB_Food:1.0* and *DB_Cuisine_per_country:1.0*), whereas the position of the anaphor can be used to determine whether it is part a semantic role or not (in our example 'it' would be part of a semantic role, i.e. *Arg1:1.0*).

We explored the added value of incorporating coreference information by including it as a separate processing step before the feature extraction. To this purpose, all reviews were processed using the COREA system trained on the complete 500,000 words subset of the SoNaR 1 corpus comprising a variety of text genres (cfr. Section 2.1.1).

Crucial for this step is that the coreference resolution is highly accurate, since an anaphor–antecedent mismatch can also lead to a semantic information mismatch. As we learned in chapter 2 that the output of COREA is not perfect, we manually annotated each 'NULL' aspect term that constitutes an anaphor–antecedent relation, in order to better asses the upper bound of incorporating coreference information for this task.

9.2.5 Experimental setup

In the framework of this subtask we explored the possible added value of coreference and semantic roles by conducting two different sets of experiments: one where the aspect terms were given and included both explicit and 'NULL' targets (Setup 1) and one where the aspect terms were derived from the previous step and included only explicit targets (Setup 2).

Setup 1: classification using gold-standard aspect terms as input

For this setup we had gold-standard aspect terms available in order to avoid error percolation from the previous step and focus on the optimization of this subtask.

First, ten-fold cross validation experiments were conducted on the development set, using LibSVM as our machine learner. We evaluated with accuracy. We refer to Section 5.2 for more details on the machine learner and evaluation metric.

As in the readability classification experiments, our main interest was in exploring whether, and if so, how the task of aspect-based sentiment analysis, which typically relies on shallow lexical characteristics, can benefit from incorporating our two semantic information layers: semantic roles and coreference. This was done in the following manner:

1. The **semantic roles** were implemented as features on top of the lexical bag-of-words token-unigram and Cornetto and DBpedia-based lexical-semantic features.
2. Contrary to the readability prediction task, **coreference resolution** was included as an additional processing step prior to classification.

To assess this latter step, the experiments on the development data were split in a setting where coreference relations were not derived beforehand and one where they were. In the latter setting, a comparison was also made between automatically-derived and gold standard coreference information to assess the true upper bound.

For the optimization of the classifier, we employed the same methodology as for the readability prediction experiments (cfr. page 84) in that we performed two different rounds of experiments:

In *Round 1*, we empirically determined which features contribute to the classification task. To this purpose we used the default hyperparameter settings of LibSVM, but with a linear kernel. The actual experiments consisted of comparing the baseline setup using bag-of-words features to a setup where the lexical-semantic and semantic role features were manually added.

In *Round 2*, we used genetic algorithms to pinpoint the optimal feature combinations in the same manner as we did for the readability prediction experiments (Section 5.2.4). Since each machine learning algorithm's performance is inherently dependent of the different parameters that are used, we performed a joint optimization in two different setups. In both setups, we allow 100 generations and set the stopping criterion to a best fitness score (accuracy) that remained the same during the last five generations. Our search starts from a population of 100 individuals and all optimization experiments are performed using the Gallop toolbox (Desmet and Hoste 2013).

1. We performed hyperparameter and feature group selection using the four feature groups we had available (i.e. bag-of-words, Cornetto, DBpedia and semantic roles). We allowed variation in LibSVM's hyperparameters as described in Section 5.2.1.
2. We performed hyperparameter selection and allowed individual feature selection among the lexical-semantic (Cornetto and DBpedia) and semantic role features. The bag-of-words features were kept together as a group, since selecting them individually would increase the search space immensely due to combinatorial explosion. Adding them as a group allows the genetic algorithm to focus its search on the remaining features. As for LibSVM's hyperparameters, we allowed the same variation as mentioned above.

In a final experiment, the optimal settings emerging from the experiments on the development data were tested on the held-out test set from which gold-standard aspect terms had also been derived.

Setup 2: classification using automatically-derived aspect terms

In this second setup we tested the fully-automatic pipeline: classifying the automatically extracted aspect terms (cfr. Section 9.1) into aspect categories. Since our aspect term extraction system only outputs explicit aspect terms, our main objective was to investigate how to derive implicit aspect terms, i.e. 'NULL' targets.

To this purpose we compared a system where coreferential anaphor–antecedent pairs were used to discover implicit aspect terms to a simple heuristic where implicit 'NULL' targets were added to our instancebase whenever a sentence was considered subjective according to our subjectivity heuristic (cfr. Section 9.1.1), but no explicit aspect terms had been detected in it.

If we reconsider the example:

(52) In La Dolce Vita we ordered a tasty [pizza], recommended by the boss. *It* tasted absolutely divine!

then the setting using coreference resolution would be able to derive that the 'It' refers back to the antecedent 'pizza' and add this anaphor as an implicit aspect term to the instancebase. The other setting, using the subjectivity heuristic, would also add this sentence in which no explicit aspect terms were indicated but where sentiment is clearly expressed to the instancebase. However, only the setting using coreference resolution would be able to derive additional lexical-semantic and semantic role features for this instance.

If we consider the following example,

(53) Wij komen zeker en vast terug!
EN: We will definitely return!

only the setting using the subjectivity heuristic would add an implicit instance for this sentence to our instancebase.

These experiments were conducted on the held-out test set. To evaluate, we decided to mimic the SemEval 2015 evaluation (Pontiki et al. 2015) and score precision, recall and F-measure because the aspect term extraction performance should also be included in this evaluation. However, unlike the (binary) task of detecting aspect terms, aspect category classification is a multiclass task which requires an averaging over the classes. We use micro-averaged F-score, which gives each instance equal weight in the evaluation. It should be noted that micro-averaged F-score is identical to accuracy if all instances in a dataset are aspect terms (i.e. there is no detection, only classification). For clarity, the aspect category classification experiments on gold-standard terms are reported as accuracy (Setup 1), but on automatically detected terms, we report micro-averaged F-score (Setup 2). The scores are directly comparable, given that accuracy is equal to micro-averaged F-score on perfectly detected terms.

Moreover, the SemEval evaluation ignores aspect categories occurring more than once in a single sentence since, in the end, we want to be able to aggregate aspect categories. For example in the sentence 'The duck, potatoes and vegetables tasted horrible', three aspect terms can be defined but these all refer to the same aspect category *Food–Quality*. In the proposed evaluation these three correctly classified instances would thus be counted as one correct aspect category classification.

9.3 Aspect polarity classification

Given a list of possible candidate terms and given that these were classified into one of the aspect categories, the final step consisted in classifying the polarity expressed towards these aspects.

For this classification task, we performed multiclass classification into one of the three possible polarity labels (positive, negative or neutral) relying on a feature space containing solely lexical features. A first prototype of this system was developed for English in the framework of SemEval 2014 Task 9 (Van Hee et al. 2014) and has also proven effective for this third subtask in SemEval 2015 Task 12 (De Clercq et al. 2015). This system was adapted to deal with Dutch text.

9.3.1 Features

For each aspect term, a variety of lexical features have been extracted based on the sentence in which the aspect term occurs.

Token and character n-gram features: binary values for each token unigram, trigram and bigram in the training data were derived, as well as character n-gram features for each character trigram and fourgram in the training data. These features are listed and illustrated in Table 9.4.

Sentiment Lexicons: we used two existing Dutch sentiment lexicons, viz. the

token1gramFeatures	e.g. *service*
token2gramFeatures	e.g. *quick service*
token3gramFeatures	e.g. *rather quick service*
character3gramFeatures	e.g. *ser–erv–rvi–vic–ice*
character4gramFeatures	e.g. *serv-ervi-rvic-vice*

Table 9.4: N-gram features together with examples.

Duoman lexicon (Jijkoun and Hofmann 2009), which was introduced in Section 9.1, and the Pattern lexicon (De Smedt and Daelemans 2012). The latter is a list of adjectives that were manually assigned a polarity value between -1 and 1 for each word sense. This list was automatically expanded using distributional extraction and synset relations. Both the Duoman and Pattern resources actually consist of two lexicons: a large list generated semi-automatically and a smaller list containing gold-standard polarity annotations.

For the feature extraction we made use of these four sentiment lexicons, namely the manually labeled subset of the Pattern lexicon, the manually annotated Duoman list, and the (semi-)automatically created Pattern and Duoman lexicons. For each lexicon, we extracted the number of positive words, the number of negative words and the number of neutral words. These three values were all averaged over sentence length. Finally, the sum of the polarity scores of all detected sentiment words was also added as a final feature (Table 9.5).

Automatic	**Gold standard**
Duoman-nrPosTokens	Duomanman-nrPosTokens
Duoman-nrNegTokens	Duomanman-nrNegTokens
Duoman-nrNeutTokens	Duomanman-nrNeutTokens
Duoman-overallValue	Duomanman-overallValue
Pattern-nrPosTokens	Patternman-nrPosTokens
Pattern-nrNegTokens	Patternman-nrNegTokens
Pattern-nrNeutTokens	Patternman-nrNeutTokens
Pattern-overallValue	Patternman-overallValue

Table 9.5: Sentiment lexicon features.

To identify the sentiment words, a lookup was performed for both the surface form and lemma of each word in the sentence and the part-of-speech needed to match the category of the corresponding sentiment word in the lexicon. To this purpose, shallow linguistic preprocessing was performed on all reviews using the LeTs Preprocess toolkit (Van de Kauter et al. 2013).

Word-shape: finally, a number of numeric and binary features were included that capture the characteristics concerning the shape of a review sentence. These features indicated whether there is character flooding or punctuation flooding present within the review sentence. This might hint at intense sentiment, e.g. 'coooooool !!!!!'. Next, we also checked whether the last token contained punctuation and count how many capitalized tokens are present within one sentence (Table 9.6). The intuition behind this last feature is, again, that capitalization might hint at sentiment, e.g. 'COOL'.

countFloodedTokens
countCapitalizedTokens
countFloodedPunctuationTokens
punctuationLastToken

Table 9.6: Word-shape features.

9.3.2 Experimental setup

For the polarity classification experiments, we again optimized on the development set and tested this optimal setting on the held-out test set.

We performed 10-fold cross validation on the development data using LibSVM and evaluated by calculating accuracy. In order to derive the optimal settings, we compared a setting with all features and the default LibSVM settings using a linear kernel to a setting where both the parameters and features were jointly optimized using the Gallop toolkit (Desmet and Hoste 2013). For these optimization experiments we relied on gold-standard aspect terms and categories, allowing us to focus on the polarity classification.

We did not perform individual feature selection on the ngram token and character features since selecting these feature groups individually, which range from 3,000 to 12,000 individual features, would increase the search space immensely due to combinatorial explosion.

For the held-out test data we performed two experiments. In the first experiment we tested the performance of our optimal system assuming gold-standard output of the two previous subtasks. In a final experiment, we tested the fully automatic pipeline, thus allowing error percolation of every step.

CHAPTER 10

Results and discussion

This chapter presents the results for the three individual subtasks. Most attention will be devoted to the second subtask of aspect category classification, since for that stage the added value of our two semantic information layers has been investigated.

10.1 Aspect term extraction

In this first step, candidate terms related to the restaurants domain have to be automatically selected from running text. As explained in section 9.1, we used the hybrid terminology extraction system TExSIS which combines linguistic and statistical information (Macken et al. 2013). To this purpose all texts were first preprocessed using the LeTs preprocessing toolkit (Van de Kauter et al. 2013), after which various words and chunks matching certain Part-of-Speech patterns were extracted as candidate terms. In the next phase of the term extraction, statistical filtering was performed. As explained, we performed additional domain-specific filtering of this TExSIS output using a subjectivity lexicon and semantic annotations based on Cornetto (Vossen et al. 2013) and DBpedia (Lehmann et al. 2013).

10.1.1 Results

Table 10.1 presents the results of the various filtering steps performed. The results are expressed in precision, recall and F-1 score. In this step we did not consider the implicit 'NULL' targets yet, the objective here was to optimize the extraction of explicit aspect terms. To this purpose the development set of 300 reviews was split in a 250 train (devtrain) and 50 test set (devtest).

	Precision	Recall	F-1
TExSIS	24.78	39.61	30.48
TExSIS + subj	29.15	**66.18**	40.47
TExSIS + subj + Cor + DB	**37.85**	59.42	**46.24**
No Heur *(TExSIS + subj + Cor + DB)*	34.77	62.32	44.64

Table 10.1: Results of the ATE experiments on the development data.

The first results (*TExSIS*) rely on the TExSIS-internal linguistic noun phrase extraction and term weighting only. Based on the term weighting filtering, each term with a value of zero was excluded from the list. In addition, terms consisting of more than six words or terms of which the part-of-speech information was less than 20% of the time a noun in the entire devtrain set, were eliminated. Finally, we preferred the largest possible units whenever there were multiple candidate term options (e.g. 'pizza margherita' over 'pizza'). It can be observed that using the TExSIS output as such leads to an F-measure of 30.48, which we consider as our baseline.

This brings us to the additional filtering steps that were performed in order to exclude candidate terms from the list. In a first step, we added subjectivity filtering (*subj*) based on the gold-standard Duoman lexicon. This step enabled us to delete the subjective adjectives in, for example, 'leuke sfeer' (EN: pleasant atmosphere) and 'vriendelijke bediening' (EN: friendly service), keeping only the base terms 'sfeer' and 'bediening'. We observe that this operation leads to an increase of over 10 F-measure points and seems especially beneficial for recall (from 39.61 to 66.18).

For the second filtering we relied on Cornetto (*Cor*) and DBpedia (*DB*). Cornetto was used to filter out all candidate terms that did not match one of six main categories in the devtrain data that (1) co-occur in the synset of the candidate term or (2) which are a hyponym/hypernym of a candidate term in the synset. For DBpedia, a list of overlapping categories was drawn up based on the devtrain data, resulting in eighteen possible DBpedia categories (cfr. Table 9.3). This was used in an attempt to find those candidate terms for which no match was found using Cornetto but that are in fact related to the restaurants domain. The results show that although this semantic filtering harms our recall, it does improve our precision a lot over the subjectivity filtering (from 29.15 to 37.85), leading to an improved F-measure of 46.24.

An additional constraint in the aspect term extraction task is that candidate terms can only be selected if sentiment is expressed towards these terms. So before we continued to the experiments on the held-out test data, we tested whether plugging in the subjectivity heuristic at the beginning of our pipeline as illustrated in Figure 9.1 and described in Section 9.1.1 is necessary. In other words, if we derive all possible aspect terms regardless of whether sentiment is expressed in a given sentence, do we get better results?

These results are presented at the bottom of Table 10.1 (*No Heur*). Though we notice that this results in a higher recall (i.e. 62.32), probably because more terms are found in sentences where the sentiment is only expressed very subtly, it also causes the precision to decrease with 3 points, leading to a precision of 34.77. The resulting F-score of 44.64 is lower than the same setting with subjectivity filtering, 46.24. These results led us to conclude that it is beneficial to perform subjectivity filtering first.

The third setting is thus our optimal setting. For the experiments on the held-out test set of 100 reviews we used this third setting, the results of which are presented in Table 10.2. In this setting, TExSIS was thus rerun on the entire development set of 400 reviews and all filtering steps were performed.

	Precision	Recall	F-1
Held-out	35.87	58.18	44.38

Table 10.2: Results of optimal ATE settings on the held-out test set.

We observe that the results on our held-out test set are lower than on our development set. Represented in absolute numbers, our held-out test set contains 373 explicit aspect term expressions or targets[1], of which 217 were found by our system, leading to the recall of 58.18%. In total, however, our system predicted 605 explicit target mentions, leading to the low precision of 35.87%. Clearly, our system would benefit from further optimization of its precision in future work. We would like to stress, however, that this evaluation is very strict in that a complete match has to be found between a predicted and gold-standard target term in order for it to be represented in the scores.

[1] Remember that targets referring to multiple aspect categories within one sentence were only counted once.

10.1.2 Error analysis

In the following, we present an error analysis that was performed on our held-out test set. Our system predicted 605 explicit aspect terms of which 217 were found to match the gold-standard explicit aspect terms (out of a total of 373 gold-standard terms). In the examples below, the wrong explicit targets are always presented between red square brackets and the correct ones matching the gold standard between green brackets.

We start by describing the problem of overgeneration. For instance, our system picked up cases of anecdotal sentences in which a reviewer describes items not related to the actual restaurant experience but that do contain sentiment words. A case in point is example 54, where our system annotated the term ‘restaurant’.

In addition, sentences in which various explicit aspect terms occur with sentiment being addressed to some but not all of these, pose a problem to our system as well, in that every possible aspect term is annotated as soon as sentiment is expressed on the sentence level. Consider 55 for an example. Here we see that three aspect terms are indicated, whereas sentiment is only expressed towards two of these. This is something that could be solved by running the third step of polarity classification. Related to this are sentences where a specific sentiment towards an aspect is expressed implicitly but other explicit terms occur in the sentence. An instance is example 56 where a negative sentiment towards the aspect *Service–general* is expressed but an explicit aspect term was labeled by our system.

(54) Als vegetariër is het allesbehalve evident een goed [restaurant] te vinden.
EN: As a vegetarian it is far from easy to find a good [restaurant].

(55) Eerst een [aperitiefje] in de mooie [tuin] en dan hebben we ons laten verrassen met heerlijke [gerechten].
EN: First a [drink] in the beautiful [garden] and then we let ourselves be surprised with the wonderful [dishes].

(56) We moesten erg lang wachten alvorens er werd afgeruimd en opgenomen voor ’t [dessert] .
EN: We had to wait really long before our table was cleaned and we could order [dessert].

Considering explicit aspect terms that were missed by our system (indicated in blue brackets), we found that sentences in which subjectivity is expressed in a more creative manner, without using explicit sentiment words (example 57), pose a problem.

(57) Vriendin had een [stoofpotje van kabeljauw] besteld daar kon je een vork in recht zetten.
EN: Girlfriend ordered a [cod stew] in which you can plant a fork straight.

Next, we have a closer look at two examples where the boundaries found for the explicit aspect term were either too small (example 33) or too large (example 34). In the latter example, we notice that our subjectivity filtering failed to filter out the subjective word 'stijfdeftige' because it was not included in the gold-standard Duoman lexicon.

(58) [Books en [Brunch]] is fantastisch!
EN: [Books en [Brunch]] is wonderful!

(59) Vlotte, doch geen [stijfdeftige[bediening]].
EN: Quick, though no [uptight [service]].

This brings us to the observation that sometimes the strict evaluation poses a problem because the boundaries are hard to define. With respect to examples such as 60, 61 and 62, it is open to discussion which boundaries are actually the correct ones. In our current evaluation setting, these were all counted as errors. This is probably too harsh, because we believe that any company which would require explicit targets to be indicated using this system would already be very pleased with this output.

(60) [Huisgemaakte [chocolademousse] als toetje] kon er nog net bij.
EN: [Homemade [chocolate mousse] as dessert] could just fit in.

(61) De [[bediening] door jonge obers] kon bovendien professioneler!
EN: The [[service] by young waiters] could have been a lot more professional!

(62) Het [uiterlijk van het [restaurant]] oogt simpel.
EN: The [look of the [restaurant]] seems simple.

To conclude, we would like to stress that the focus of this evaluation was on the extraction of explicit aspect terms, ignoring the implicit aspect terms, i.e. the 'NULL' targets. We hypothesize that coreference information is useful for pinpointing implicit aspect terms and can be incorporated immediately after the actual term extraction. As the incorporation of this additional information source can only be evaluated after the aspect terms have been assigned to specific categories, the added value of this step will be investigated in the next subtask, i.e. aspect category classification.

10.2 Aspect category classification

In this subtask, aspect terms have to be classified into broader aspect categories. This refers to detecting both the six main categories (e.g. *Food*) and the various attributes or subcategories (e.g. *Prices*) in one go. Within this subtask we were able to explore the possible added value of our two semantic information layers: coreference and semantic roles. To this purpose, we conducted two different sets of experiments:

In Setup 1 (Section 10.2.2), we assume the aspect terms are given, i.e. gold standard, which means we have both explicit and implicit ('NULL') aspect terms available and we can really focus on the optimization of the aspect category classification. We explore the added value of incorporating more semantic features into the feature space in the form of lexical-semantic and semantic role features. In this respect, we also investigate the added value of resolving coreference prior to the classification step. This was done by performing the optimization experiments in a setting where coreferential anaphor–antecedent pairs were not derived beforehand and one where they were.

In Setup 2 (Section 10.2.2), we assume a fully-automatic pipeline in that we start from the explicit aspect terms that came out of the previous step of aspect term extraction. In this setup our focus is on discovering implicit ('NULL') aspect terms. We compare a setting where implicit aspect terms are found using gold-standard coreference resolution and one where a simple subjectivity heuristic is used to this purpose.

10.2.1 Setup 1: classification using gold-standard aspect terms as input

Optimization on the development data

As explained in the experimental setup, we conducted optimization experiments on the development data in (i) a setting where coreferential anaphor–antecedent pairs were not derived beforehand and (ii) one where they were. In the latter setting, both gold-standard and automatically-derived coreference relations were used in order to investigate the true upper bound of incorporating this type of information. All experiments were performed using 10-fold cross-validation evaluated with classification accuracy.

In Table 10.3, we present the results (expressed in accuracy) of the experiments without coreference (i). The main purpose of these experiments was to explore the added value of adding semantic information in the form of lexical-semantic (*lexsem*) and semantic role features (*srl*), separately or together (*lexsem + srl*), to a baseline setup where only bag-of-words token unigrams (*bow*) were used for the classification. Since coreference resolution was not performed prior to classificiation, only the explicit aspect terms could benefit from this.

For the optimization of the classifier, two different rounds of experiments were performed: one where the added value of semantics was empirically verified by gradually adding more features (*Round 1*) and one where genetic algorithms were used to this purpose (*Round 2*). In the second round, a distinction was made between jointly optimizing the hyperparameters and feature groups and jointly optimizing LibSVM's hyperparameters and the individual semantic features. For our baseline, consisting solely of the bag-of-words features, only hyperparameter optimization was performed in the second round.

	Round 1	*Round 2*	
bow	53.28	54.69	
		Joint optimization	
		featgroups	indfeats
bow + lexsem	**60.72**	**62.94**	63.16
bow + srl	54.80	56.16	56.70
bow + lexsem + srl	60.01	62.89	**63.27**

Table 10.3: (i) Results of cross-validation experiments on the development data without performing coreference resolution prior to classification.

Round 1: We observe that both semantic information sources improve the performance when compared to the baseline in different gradations. Whereas the semantic role features allow for a mild improvement of 1.47 points, the lexical-semantic Cornetto and DBpedia features allow for an improvement of 7 points. We also notice that when we combine both semantic information sources (*lexsem+srl*) with the bag-of-words features this leads to an improvement over the baseline of 6.73 points, which means that this setting does not outperform the result achieved when only incorporating lexical-semantic information (an accuracy of 60.01 versus one of 60.72).

Round 2: Overall, we notice that all setups benefit from jointly optimizing the hyperparameters and features. We go from a best score of 60.72 using the default settings and only the *lexsem* features to one of 63.27 where both the hyperparameters and all semantic features have been optimized individually. If we compare both optimization setups, viz. jointly optimizing feature groups versus jointly optimizing the individual semantic features, we observe that the best results are achieved with the individual feature selection experiments. In the best setup, both lexical-semantic and semantic role features are included, resulting in an accuracy of 63.27. This means that both semantic information sources contribute to the task, though the added value of the lexical-semantic features is much more outspoken.

In a next setting (ii) coreference resolution was included as an additional processing step prior to classification. Having coreference information available should allow us to derive additional semantic information for those 'NULL' targets constituting an anaphor–antecedent pair. We differentiate between a setup where we incorporate this information assuming we have a perfect coreference resolution system (COREF GOLD), i.e. using gold-standard coreferential links, and a setup where coreference relations are resolved automatically (COREF AUTO). This should allow us to truly explore the possible added value of incorporating coreference resolution as an additional processing step prior to classification.

The results, expressed in accuracy, are presented in Table 10.4. To facilitate comparison, the results of the previous experiments (without coreference) are added in grey. The best individual results are indicated in bold.

	Round 1		*Round 2*			
bow	53.28		54.69			
			Joint optimization			
			featgroups		indfeats	
COREF GOLD						
bow + lexsem	60.72	**61.26**	**62.94**	62.78	63.16	**63.59**
bow + srl	54.80	54.80	56.16	56.16	56.70	56.59
bow + lexsem + srl	60.01	60.99	62.89	62.67	63.27	62.34
COREF AUTO						
bow + lexsem	**60.72**	59.63	**62.94**	60.77	**63.16**	60.88
bow + srl	54.80	54.42	56.16	55.89	56.70	56.76
bow + lexsem + srl	60.01	59.36	62.89	60.77	63.27	60.61

Table 10.4: (ii) Results of cross-validation experiments on the development data with performing coreference resolution prior to classification.

Round 1: When using gold standard coreference information, we observe that the results increase (*bow+lexsem* and *bow+lexsem+srl*) or remain unchanged (*bow+srl*). As in the previous experiments, the best score is achieved with the lexical-semantic features alone, viz. an accuracy of 61.26. This indicates that including coreferential links between anaphor–antecedent pairs is beneficial. If we resolve coreference automatically, however, we see that all our results decrease when compared to the results we achieved without adding coreference.

Round 2: Again, we notice that all setups benefit from joint optimization. Using gold standard coreference information, we see that the best overall result is achieved when using gold standard coreference information and lexical-semantic features in the form of Cornetto and DBpedia features that have been individually optimized. This is the best overall score that is achieved for this subtask, i.e. an accuracy of 63.59. Again, we notice that, using an automatic coreference resolver to this purpose, the results almost always deteriorate compared to the results where no coreference information was included (in grey).

Analysis of the optimal settings

From the experimental results, we can conclude that the lexical-semantic features contribute more to the aspect category classification task than the semantic role features. Although in the setting without coreference the best result is achieved using both semantic information sources, this modest added value of the semantic role features disappears completely in the best setting using (gold standard) coreference information.

In order to gain more insight in the optimal settings and before testing these on our held-out test set, we will now briefly discuss which hyperparameters and features were selected in the best setup without (Table 10.3) and with (Table 10.4) coreference resolution:

(a) The best accuracy achieved in the setting without coreference was 63.27, after jointly optimizing the hyperparameters and *bow + lexsem + srl* features.

(b) The best accuracy achieved in the setting with gold standard coreference was 63.59, after jointly optimizing the hyperparameters and *bow + lexsem* features.

Since, in both settings, the best results were reached using joint optimization with individual feature selection, our discussion is limited to these two setups. In order to select the optimal hyperparameters and features, we started from the *k*-nearest fitness solution set; these are the individuals that obtained one of the top *k* fitness scores, given an arithmetic precision. Following Desmet (2014), we used a precision of four significant figures and set *k* to three.[2]

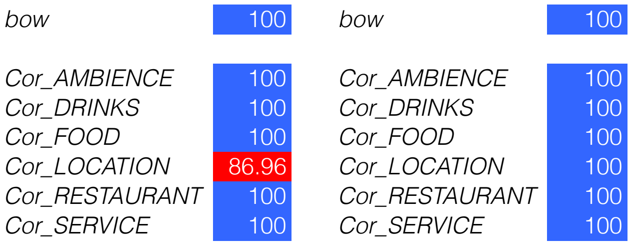
Overall, the (a) experiments required 19 generations to reach the optimal setting, whereas the (b) experiments required 15. Considering LibSVM's hyperparameters that were selected, we found that in both optimal setups, RBF kernels

[2] This was explained in Section 6.2.2.

are preferred over simpler linear kernels, and a high cost value of 2^{10} is selected, together with a very low gamma parameter (2^{-14}).

We also had a closer look at the exact features that were selected in both setups. As was also done for the readability prediction experiments, the importance of the features is visualized using a color range: the closer to blue, the more the feature in question was turned on, and the closer to red, the less important the feature was for reaching the optimal solution. The numbers within the cells represent the same information but then percentagewise.

Figure 10.1 clearly illustrates that the bag-of-words and all Cornetto features are crucial in both settings.



bow	100
Cor_AMBIENCE	100
Cor_DRINKS	100
Cor_FOOD	100
Cor_LOCATION	86.96
Cor_RESTAURANT	100
Cor_SERVICE	100

(a)

bow	100
Cor_AMBIENCE	100
Cor_DRINKS	100
Cor_FOOD	100
Cor_LOCATION	100
Cor_RESTAURANT	100
Cor_SERVICE	100

(b)

Figure 10.1: Selected bag-of-words and Cornetto features in the optimal experiments without (a) and with (b) coreference.

For the DBpedia features, listed in Figure 10.2, at first sight there seems to be a difference between the features that are turned on in the optimal settings. However, on a closer look, it turns out that twelve of the sixteen features get the same value and, that in total, twelve features are turned on in both settings.

Regarding the semantic role features, only the setup where coreference was not included as an additional processing step prior to classification (a), achieved better results when these features were included. Figure 10.3 illustrates which features were considered important. Contrary to our expectations, the predicate tokens do not seem to contribute to the overall classification in that they do not allow for the assignment of more fine-grained main-attribute category labels. As for the actual semantic roles, we observe that only the two highest arguments, i.e. the Arg3 and Arg4 features, are found to be good predictors, together with five from the twelve possible modifier features.

DB_Alcoholic_beverages	100
DB_Bread_and_pastry	100
DB_Breadspread	52.17
DB_Candy	52.17
DB_Catering_industry	100
DB_Cattle	4.35
DB_Cuisine_per_country	65.22
DB_Dish	13.04
DB_Edible_plant	47.83
DB_Fish	34.78
DB_Food	69.56
DB_Food_terminology	82.61
DB_Furniture	100
DB_Herbs_and_spices	100
DB_Gastronomy	65.22
DB_Not-alcoholic_beverages	52.17
DB_Nutrition	47.83
DB_Poultry	0

(a)

DB_Alcoholic_beverages	100
DB_Bread_and_pastry	59.091
DB_Breadspread	100
DB_Candy	100
DB_Catering_industry	0
DB_Cattle	77.27
DB_Cuisine_per_country	100
DB_Dish	0
DB_Edible_plant	0
DB_Fish	50
DB_Food	63.64
DB_Food_terminology	18.18
DB_Furniture	72.73
DB_Gastronomy	0
DB_Herbs_and_spices	100
DB_Non-alcoholic_beverages	0
DB_Nutrition	100
DB_Poultry	95.45

(b)

Figure 10.2: Selected DBpedia features in the optimal experiments without (a) and with (b) coreference.

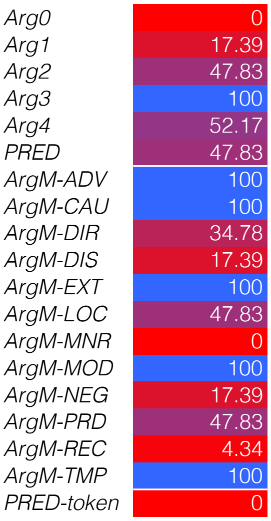
Testing the optimal settings on the held-out test set

Having determined the optimal hyperparameters and features, the two optimal results – setup (a) and setup (b) – were used to train two models on the entire development data and test these models on our held-out test set in which we also assumed perfect aspect terms.

In both settings, we achieved an accuracy of **66.42**, which is three points higher than the best accuracy scores on our development set using 10-fold cross validation. This also indicates that on our held-out test set there is no difference between the accuracy obtained with and without performing (gold standard) coreference resolution prior to classification.

A closer look at coreference

From the above-mentioned results, it can be concluded that the added value of including coreference information as an additional step before feature extraction



(a)

Figure 10.3: Selected semantic role features in the optimal experiment without coreference (a).

is not outspoken. Using the output from the COREA tool led to an overall decrease in performance, the main reason being that the COREA output is not accurate enough. As a consequence, wrong antecedents have been linked to anaphors, causing faulty lexical-semantic features, or wrong anaphors were indicated, leading to faulty semantic role features.

However, our results also showed that incorporating these as gold standard anaphor–antecedent relations leads to the best overall score achieved for the subtask of aspect category classification, i.e. an accuracy of 63.59 after jointly optimizing LibSVM's hyperparameters and performing individual feature selection. If we compare this best score to the best individual score achieved in the setting without coreference, we see that the difference is marginal, however i.e. 63.27 without versus 63.59 with coreference.

In order to understand this better, we had a closer look at all the implicit aspect terms present in our development set and the gold-standard coreference relations. In total, our development dataset contains 563 'NULL' targets (which corresponds to almost 30% of all aspect term expressions). The results of this manual analysis are visualized in Figure 10.4.

On a total of 563 'NULL' targets, 417 constitute a truly implicit target, such as

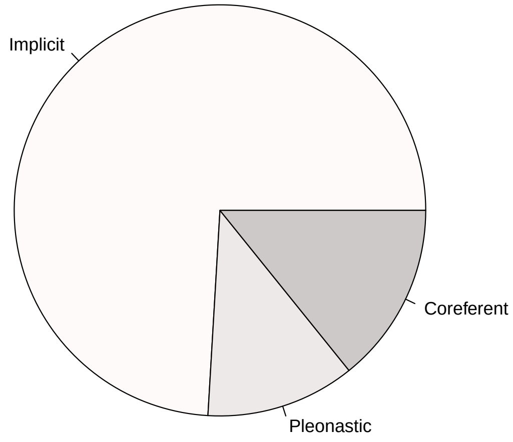


Figure 10.4: Division of the 'NULL' targets in the development dataset.

example 63 where a negative opinion is clearly expressed towards the *Service*. In addition, 66 'NULL' targets are pleonastic pronouns[3], e.g. example 64, where the anaphor 'ze' is used to refer to the *Service* but the antecedent is nowhere to be found in the remainder of the review.

(63) Niets hartelijk, vriendelijk!
EN: Neither cordial nor friendly!

(64) Als je een goedkopere fles bestelt geven *ze* een onvoorstelbaar denigrerende blik.
EN: If you order a cheaper bottle *they* give you this horrible condescending look.

This leaves us with 80 'NULL' targets that are truly coreferent and constitute an anaphor–antecedent pair. As a consequence, using perfect coreference information allowed us to derive additional semantic information for circa 5% of our instancebase. This explains the small performance increase when compared to our best setting when coreference was not included, from an accuracy of 63.27 to one of 63.59. The results on our held-out test set even indicated there is no difference.

[3] We are aware of the potentially terminological ambiguity when referring to these pronouns as pleonastic. This term is normally used for an empty or dummy pronoun, as in the prototypical example 'It snows'. In our case some of these pronouns are actually referential, but their referents are not explicitly referred to in the text.

10.2.2 Setup 2: classification using automatically-derived aspect terms

In this second setup, we tested the fully-automatic pipeline. These experiments were conducted on the held-out test set using the optimal settings from the previous setup (viz. best setup (a) and (b) as presented on page 167) trained on the entire development data.

First, we evaluate the aspect category classification relying solely on the explicit aspect terms. Next, the main challenge was to investigate how to derive implicit aspect terms, i.e. 'NULL' targets. To this purpose, we conducted three experiments:

1. In a first experiment, gold-standard coreferential anaphor–antecedent pairs were used to discover additional implicit aspect terms on top of the explicit aspects. For the aspect category classification, this also implied that for some of those implicit aspect terms additional semantic information, in the form of lexical-semantic features, was derived.
2. In a second experiment, implicit aspect terms were derived by relying on our subjectivity heuristic. If a sentence occurred in which sentiment was clearly expressed, but no explicit aspect terms had been identified, an implicit aspect term was generated. For this step, the feature space of the additional implicit terms relies solely on bag-of-words features.
3. In the third and final experiment, we combined both the gold-standard coreference resolution and subjectivity heuristic to uncover implicit aspect terms. In this final step, for some implicit aspect terms, i.e. those constituting an anaphor–antecedent relation, we could also derive additional lexical-semantic features.

It should be noted that we evaluated the performance in this setup by calculating micro-averaged precision, recall and F-1, because the term extraction should also be taken into account.[4] The results are presented in Table 10.5.

We observe that, compared to a setup where no implicit aspect terms are added, adding implicit coreferential links to our system leads to a mild improvement, from an F-measure of 47.96 to one of 48.28, mainly because more implicit aspect terms are found (improved recall). Using the subjectivity heuristic to this purpose, however, is clearly the best strategy and leads to an F-measure of 54.10.

[4] We refer to page 156 for a discussion on the compatibility of accuracy and micro-averaged F1.

	Precision	Recall	F-1
Only explicit aspect terms	53.70	43.33	47.96
Experiment 1	52.93	44.38	48.28
Experiment 2	**53.82**	**54.39**	**54.10**
Experiment 3	53.47	54.04	53.75

Table 10.5: Results of aspect category classification on explicit aspect terms and of the three experiments where implicit aspect terms were added.

In the final experiment, where we combine both techniques to derive implicit aspect terms, under the hypothesis that additional semantic features for the implicit anaphor–antecedent pairs might result in better performance, we observe a slight drop in performance.

A closer look at coreference

From the results of our Setup 2 experiments, we can conclude that, apparently, a simple subjectivity heuristic outperforms a system where gold-standard coreferential anaphor–antecedent pairs are added.

Nevertheless, we were surprised that the final setting combining both (Experiment 3) did not outperform the setting where only the subjectivity heuristic was used (Experiment 2), mainly because in the latter setting our system could rely on additional semantic information besides bag-of-words features to classify the implicit aspect terms.

If we have a closer look at the anaphor–antecedent pairs in our held-out test set, however, we found that these 100 reviews contain 210 implicit aspect terms, of which 154 are truly implicit, 31 are pleonastic[5] and only 15 are truly coreferential.

In Experiment 1, these 15 instances were thus added to the instancebase, but this did not necessarily mean they were all classified correctly, hence the slight increase in recall but not in precision.

When comparing the instancebase of our held-out test set of Experiment 2 with the one of Experiment 3, we observed that, in total, the feature vectors of only five implicit aspect terms had received additional semantic information in the form of lexical-semantic features. In three of these cases, this led to a different

[5] See page 171 for a brief explanation with regard to the use of the term pleonastic in this context.

aspect category prediction, as illustrated in the examples below.

(65) ...[restaurant]... *Er* heerst een optimale gezelligheid, kalmte en alles wordt niet geforceerd en op het gemak gedaan zonder laks te lijken.
EN: ...[restaurant]... *There* reigns an optimal coziness, calmness and everything is done at ease without appearing sloppy.
- Prediction experiment 2 (only lexical): *Service–General*
- Prediction experiment 3 (lexical and semantic): *Service–General*

(66) ...[café]... En het is *hier* veel te duur.
EN: ...[pub]... And it is way overpriced *here*.
- Prediction experiment 2 (only lexical): *Food–Prices*
- Prediction experiment 3 (lexical and semantic): *Food–Prices*

(67) ...[vrouw]... Geen moment is *ze* gehaast over gekomen en nam voor iedereen de tijd.
EN: ...[woman]... Not a single moment did *she* appear rushed and she took time for everyone.
- Prediction experiment 2 (only lexical): *Service–General*
- Prediction experiment 3 (lexical and semantic): *Food–Style&Options*

(68) ...[eten]... Ik vind het ook wat duur voor *wat* we hebben gekregen.
EN: ...[food]... I find it a bit overpriced for *what* we actually received.
- Prediction experiment 2 (only lexical): *Food–Prices*
- Prediction experiment 3 (lexical and semantic): *Food–Quality*

(69) ...[mosselen]... Voor minder dan 20 euro krijg je *ze* niet.
EN: ...[mussels]... For less than 20 euros you will not get *them*.
- Prediction experiment 2 (only lexical): *Service–General*
- Prediction experiment 3 (lexical and semantic): *Food–Quality*

In examples 65 and 66, we observe that the prediction remains the same. These are actually wrong predictions (hence their red colour) as the gold-standard aspect categories for these 'NULL' aspect are *Restaurant–Ambience* and *Restaurant–Prices*, respectively.

In examples 67 and 68, we notice that our classifier predicts a different, i.e. wrong, category in experiment 3 where semantic features were derived for these instances on top of the lexical ones. This is because for these two instances, the additional lexical-semantic information derived on the basis of Cornetto, led to a different interplay, and thus caused our classifier to predict these faulty categories.

In example 69, finally, we observe that a wrong aspect category is predicted in both experiments. However, in experiment 3, the prediction is closer to the gold-standard category, *Food–Prices*, mainly because the antecedent 'mosselen' (mussels) received the DBpedia feature: DB_Food.

10.3 Aspect term polarity classification

Given a list of possible candidate terms and given that these were classified into one of the aspect categories, the final step consisted in classifying the polarity expressed towards these aspects.

For these experiments, we again first optimized on the development data, after which this optimal setting was tested on the held-out test data.

We performed 10-fold cross validation experiments on the development data using LibSVM and evaluated by calculating accuracy. In order to derive the optimal settings, we compared a setting with all features and the default LibSVM settings using a linear kernel to a setting where both the parameters and features were jointly optimized. For these optimization experiments, we relied on gold-standard aspect terms and categories. The results of these optimization experiments are presented in Table 10.6.

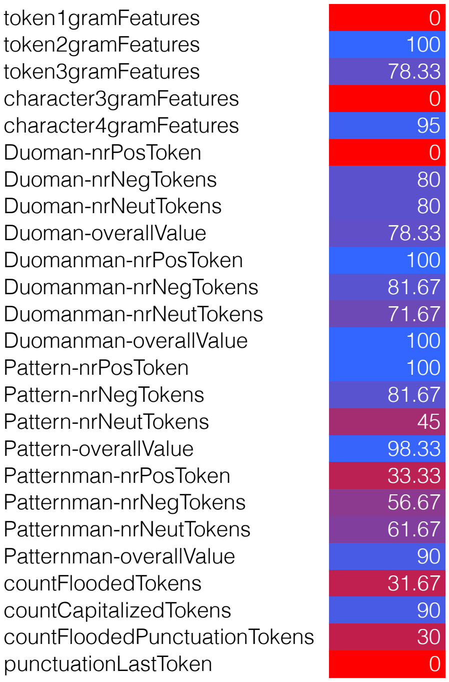
	Default	Joint optimization
All features	76.40	**79.06**

Table 10.6: Results of the optimization experiments on the development set.

We observe that our system using only lexical features benefits from joint optimization, as did all the other classification experiments in this dissertation, and goes from an accuracy of 76.40 to one of 79.06.

We have a closer look at these optimal settings, using the same k-fitness evaluation as explained above. Considering the hyperparameters, we observe that our system prefers a linear kernel with a cost-value of 2^{-4}.

Figure 10.5 presents which features are considered important after optimization, using the blue-red colour gradations. Considering the n-gram features, we observe that the bigram and trigram token features and the character fourgram features are important for achieving the optimal result. From both the Duoman and Pattern lexicon features, six features are retained. Overall, we observe that the Duoman lexicon features are turned on more often. Finally, if we have a closer look at the word-shape features, it appears that only the number of capitalized tokens present in a sentence comprising an aspect term is selected in the optimal setting.



token1gramFeatures	0
token2gramFeatures	100
token3gramFeatures	78.33
character3gramFeatures	0
character4gramFeatures	95
Duoman-nrPosToken	0
Duoman-nrNegTokens	80
Duoman-nrNeutTokens	80
Duoman-overallValue	78.33
Duomanman-nrPosToken	100
Duomanman-nrNegTokens	81.67
Duomanman-nrNeutTokens	71.67
Duomanman-overallValue	100
Pattern-nrPosToken	100
Pattern-nrNegTokens	81.67
Pattern-nrNeutTokens	45
Pattern-overallValue	98.33
Patternman-nrPosToken	33.33
Patternman-nrNegTokens	56.67
Patternman-nrNeutTokens	61.67
Patternman-overallValue	90
countFloodedTokens	31.67
countCapitalizedTokens	90
countFloodedPunctuationTokens	30
punctuationLastToken	0

Figure 10.5: Selected features in the optimal polarity experiments.

For the experiments on our held-out test data, we first trained a model on our entire development set using these optimal settings and tested it on our held-out test set where we assume gold-standard aspect terms and categories. This results in an accuracy of **81.23**.

In a final experiment, we tested the fully automatic pipeline, thus allowing error percolation of every step. For this, we tested our optimal trained model on the best setting where both explicit and implicit aspect terms had been derived (Experiment 3, which reached an F-measure of 54.10). On this output we achieve a polarity classification accuracy of 39.70, which underlines once more the high error percolation of the previous steps.

Error analysis

In order to get some insights into what goes wrong with the polarity classification, we performed an error analysis on the output of our optimized system on the held-out test set while assuming gold-standard aspect categories. For this experiment we reached an accuracy of 81.23.

In absolute numbers, of the in total 602 explicit aspect terms towards which sentiment is expressed, in total 489 instances were correctly classified (156 negative and 333 positive). Our system did learn to label the neutral label, but only assigned this label once in our held-out test set, which turned out to be a wrong prediction. When neutral labels had to be predicted, we observe that our system classified these mostly as positive.

Overall, our system has a slight bias towards the positive class, which can be explained by the fact that in our dataset overall more positive opinions are expressed (cfr. Section 8.3) and that we optimized on accuracy thus giving equal weight to each class label. If we would use majority voting and predict every possible target as having a positive polarity, this would already result in an accuracy of 67.28.

If we consider the following example sentence:

(70) De [pizza] was heerlijk maar de [service] trok op niks.
EN: The [pizza] was delicious but the [service] left much to be desired for.

our current system will make at least one wrong classification when assigning the polarity because it works on lexical information derived from the entire sentence in which an aspect term occurs, leading to exactly the same feature vectors for both instances. In our held-out test set, however, we found that, in total, of the 262 targets occurring together with another target in one specific sentence, only 15 (7.63%) constituted a different polarity. This explains our overall good performance on this particular dataset.

To conclude, we would like to stress that this subtask was not our main focus, which is why we did not further optimize our polarity classifier by making some straightforward extensions, such as adding negation or modality cues, use dependency relations to limit the context around the aspect terms or use specific opinion relations to this purpose, etc. Instead, our focus was on the added value of incorporating deep semantic information in the pipeline as explained extensively in the previous sections. Nevertheless, we have shown that using a classifier relying solely on lexical features already achieves satisfying results for the subtask of aspect polarity classification. These results would, of course, have to be corroborated on larger and different datasets.

10.4 Conclusion

In this second part, we incorporated our two deep semantic processing techniques in an aspect-based sentiment analysis pipeline.

Since no benchmark datasets existed for Dutch, we described how we collected and annotated a corpus of restaurant reviews that could serve as training and evaluation data for this task. To this purpose, we adapted established guidelines for this task to Dutch. We described that our dataset contains both explicit and implicit aspect terms.

We have explained how we built an aspect-based sentiment analysis system which consists of three individual subtasks: aspect term extraction, aspect category classification and aspect polarity classification. For each of these subtasks, we optimized the performance on a development set and tested this optimal setting on a held-out test set.

For the first step of aspect term extraction, we investigated to what extent an existing end-to-end terminology extraction system could be applied to this task. We used a reduced version of TExSIS and performed additional domain-specific filtering allowing us to optimize the extraction of candidate terms. We explained how this step could benefit from additional adaptations with a specific focus on precision and that the evaluation used was probably too strict.

The main focus of this part was on the second subtask, i.e. aspect category classification since this is where we investigated the added value of our two semantic information layers. We explained how we performed two different experimental setups, one where we assumed gold-standard aspect terms (Setup 1) and one where we relied on the output of the previous subtask (Setup 2).

We explained how in Setup 1 we focussed on investigating the added value of performing coreference resolution prior to classification in order to derive which implicit aspect terms (anaphors) could be linked to which explicit aspect terms (antecedents). In order to determine both the upper bound and actual contribution of adding coreference information, we made a difference between gold-standard and automatically derived anaphor–antecedent pairs. In these experiments, we explored how the performance of a baseline classifier relying on lexical information alone would benefit from additional semantic information in the form of lexical-semantic and semantic role features. We hypothesized that if coreference resolution was performed prior to classification, more of this semantic information could be derived, i.e. for the implicit aspect terms, which would lead to a better performance. Our results, however, revealed a very moderate performance gain. Also when comparing the semantic role features to the lexical-semantic features, it seemed that especially the latter features allow for a better performance.

In Setup 2, we investigated how to resolve implicit aspect terms. We compared a setting where gold-standard coreference resolution was used to this purpose to a setting where the implicit aspects were derived based on a simple subjectivity heuristic. Our results revealed that using this heuristic results in a better coverage and performance. We would like to stress, however, that these findings need to be corroborated on larger datasets.

For the final subtask of aspect polarity classification, we explained how we adapted an existing English system to deal with Dutch text. We have shown that this basic system, which relies solely on lexical information, already yields satisfying results.

CHAPTER 11

Conclusion

In this thesis, we set out to explore the added value of incorporating deep semantic processing in a readability prediction system and an aspect-based sentiment analysis pipeline. Our focus was on coreference resolution and semantic role labeling. Many systems and resources have been developed for these semantic processing techniques in the past, even for relatively under-resourced languages such as Dutch. Nevertheless, their added value in end-user applications has not sufficiently been examined (Poesio et al. 2010, Màrquez et al. 2008).

We opted for the task of readability prediction because the existing systems traditionally rely on more superficial text characteristics, thus leaving room for improvement. Although more complex linguistic features trained on various levels of complexity have proven quite successful when implemented in a readability prediction system (Pitler and Nenkova 2008, Kate et al. 2010, Feng et al. 2010), there is still no consensus on which features are actually the best predictors of readability. As a consequence, when institutions, companies or researchers in a variety of disciplines wish to use readability prediction techniques, they still rely on the more outdated superficial characteristics and formulas (van Boom 2014).

The same goes for the task of aspect-based sentiment analysis, which is a very fine-grained sentiment analysis task within the relatively new field of opinion mining in NLP. State-of-the-art systems developed for recent aspect-based sentiment analysis challenges (Pontiki et al. 2014), rely almost exclusively on lexical features and although the added value of coreference resolution is described as crucial in survey works (Liu 2012, Feldman2013), qualitative research in this direction is scarce.

The main contribution of this thesis is that we investigated the added value of deep semantic processing in the form of coreference and semantic role information for these two specific tasks. There were no datasets or systems available. Therefore, we collected and annotated data for both tasks and developed a state-of-the-art Dutch classification-based readability prediction system and a first end-to-end Dutch aspect-based sentiment analysis pipeline. The datasets and systems developed within this PhD research will undoubtedly be useful for future research.

A range of experiments was conducted to answer each of our three specific research questions (see Section 1.3 or below). In this chapter, we summarize the main findings, discuss the limitations of our research and list some prospects for future research.

11.1 Deep semantic processing

RQ 1: How robust are coreference resolution and semantic role labeling systems when applied to a large variety of text genres?

In order to investigate this, we adapted an existing Dutch coreference resolution system (COREA) and developed a first classification-based semantic role labeler (SSRL) for Dutch trained on a substantial amount of data. As our two end-user applications would require these systems to work with non-newspaper text material, we tested the cross-genre portability of COREA and SSRL using a large corpus of semantically annotated data comprising a variety of genres, SoNaR 1.

On the basis of these cross-genre experiments, we concluded that for both the coreference resolution and the semantic role labeling system to be robust, it is advisable to train on a substantial amount of training data comprising various text genres. If possible, it is best to also include a (small) amount of genre-specific training data.

We therefore retrained our systems on a variety of text material that had been manually annotated with both coreference and semantic role information in the framework of the SoNaR 1 corpus, comprising six text genres: administrative texts, autocues, texts used for external communication, instructive texts, journalistic texts and wikipedia texts.

11.2 Readability prediction

RQ 2: Can we push the state of the art in generic readability prediction by incorporating deep semantic text characteristics in the form of coreference and semantic role features?

11.2.1 Readability prediction system

For the construction of a readability prediction system using supervised machine learning, three steps can be roughly distinguished. First of all, a readability corpus containing text material of which the readability will be assessed must be composed. Second, a methodology to acquire readability assessments has to be defined. Finally, based on the readability corpus and the acquired assessments, prediction tasks can be performed.

Traditionally, a readability corpus consists of reading material for language learners and is assessed by experts indicating grade levels or absolute scores. There is a lack of general domain corpora, especially for Dutch, and other methodologies for assessing readability are scarce. We described in close detail how such a general-purpose corpus consisting of a large variety of text material was built and how this corpus was assessed for readability. In this respect, we proposed a new assessment technique which had not been used in readability assessment before, namely crowdsourcing, which we have shown to be a viable alternative to using expert labels for assessing readability.

Regarding the actual readability prediction system, we explained which new features, viz. five coreference and twenty semantic roles features, were implemented together with other state-of-the-art features encoding traditional, lexical, syntactic and other semantic information. We defined two classification tasks for readability prediction: a binary and a multiclass task, each of which involved the comparison of text pairs.

11.2.2 Added value

We explored the added value of the new features derived from our two semantic layers by performing two different rounds of experiments. In the first round these features were manually in- or excluded and in the second round joint optimization experiments were performed using a wrapper-based feature selection system based on genetic algorithms. In both setups, we investigated whether there was a difference in performance when these features were derived from gold-standard or automatically-generated information, which allowed us to assess the true upper bound of incorporating this type of information.

Our results revealed that readability classification definitely benefits from the incorporation of semantic information in the form of coreference and semantic role features. The best results for both tasks were achieved after jointly optimizing the hyperparameters and semantic features using genetic algorithms. Contrary to our expectations, we observed that the upper bound of our system was achieved when relying on the automatically predicted deep semantic features. This is an interesting result, because in the end we want to be able to predict readability based exclusively on automatically-derived information sources. In the work performed by Feng et al. (2010) the added value of features derived from coreference information was not corroborated due to the high level of errors produced by the coreference system. We can conclude that for our study the features derived from automatically predicted coreference resolution provide an added value and thus push the state of the art.

11.2.3 Limitations and future work

Our main focus was on incorporating deep semantic information in the form of coreference and semantic roles. Though our results have revealed that features derived from these provide an added value, we believe additional research should be performed on different datasets in order to further corroborate these results. In this respect, we would also like to stress that other deep linguistic processing techniques still need to be explored, too. In addition, it will be interesting to explore the interplay between syntax and semantics, semantic roles versus dependency syntax, in closer detail in future research.

In the framework of this dissertation we did not have the resources available to test our system on new unseen texts, which is something which will be very interesting to investigate and is the next logical step. As an alternative for approaching readability prediction as a classification task, we are strongly convinced that our corpus and methodology can also be translated into a regression task where scores are predicted for an individual text instead of classifying two texts. If a teacher, for example, wishes to select text material with a certain difficulty, a score might be handy. We have already performed experiments in this direction which led to promising results (De Clercq and Hoste under review).

Though one of our main concerns was to build a generic system that is capable of assessing the readability of texts we are all confronted with on a daily basis, we believe it will also be interesting to adapt this system to work on different, more specific, domains. First contacts have been established to test our current system on Dutch legal texts and texts for second language learners.

11.3 Aspect-based sentiment analysis

RQ 3: Does more information on discourse entities and their roles help to pinpoint the different agents and aspects in aspect-based sentiment mining?

11.3.1 Aspect-based sentiment analysis pipeline

We developed the first aspect-based sentiment analysis system for Dutch. To this purpose we collected a corpus of Dutch restaurant reviews and annotated each review with aspect term expressions and sentiment, on the basis of guidelines that were developed for a similar English task but that we adapted to Dutch. For the creation of our system, we distinguished three individual subtasks: aspect term extraction, aspect category classification and aspect polarity classification. For each of these subtasks, we optimized the performance on a development set and tested this optimal setting on a held-out test set.

For the first step of aspect term extraction, we investigated to what extent an existing end-to-end terminology extraction system could be applied to this task. We used a reduced version of TExSIS and performed additional domain-specific filtering allowing us to optimize the extraction of candidate terms. We explained how this step could benefit from additional adaptations with a specific focus on precision and that the evaluation used, which takes only exact matches into account, was probably too strict.

The main focus of this part was on the second subtask, i.e. aspect category classification, since this is where we investigated the added value of our two semantic information layers. We designed two different experimental setups, one where we assumed gold-standard aspect terms (Setup 1) and one where we relied on the output of the previous subtask (Setup 2).

For the final subtask of aspect polarity classification, we explained how we adapted an existing English system to deal with Dutch text. We have shown that this basic system, which relies solely on lexical information, already yields satisfying results when jointly optimized using genetic algorithms.

11.3.2 Added value

As mentioned above, the added value of our two semantic information layers was investigated in the framework of the second subtask of aspect category classification.

In Setup 1 we focussed on investigating the added value of performing coreference resolution prior to classification in order to derive which implicit aspect terms (anaphors) could be linked to which explicit aspect terms (antecedents). We made a difference between gold-standard and automatically-derived anaphor–antecedent pairs. In these experiments, we explored how the performance of a baseline classifier relying on lexical information alone would benefit from additional semantic information in the form of lexical-semantic and semantic role features. We hypothesized that if coreference resolution was performed prior to classification, more of this semantic information could be derived, i.e. for the implicit aspect terms, which would result in a better performance. In this respect we optimized our classifier using a wrapper-based approach to feature selection and we compared a setting where we relied on gold-standard anaphor–antecedent pairs to a setting where these had been predicted.

Our results revealed a very moderate performance gain and underlined that incorporating coreference information only proves useful when it is perfect. When coreference relations were derived automatically, this led to an overall decrease in performance because of semantic mismatches between implicit anaphors and explicit antecedents. When comparing the semantic role to the lexical-semantic features, it seemed that especially the latter features allow for a better performance.

In Setup 2, we investigated how to resolve implicit aspect terms. We compared a setting where gold-standard coreference resolution was used to this purpose to a setting where the implicit aspects were derived based on a simple subjectivity heuristic. Our results revealed that using this heuristic results in a better coverage and performance, which means that, overall, it was difficult to find an added value in resolving coreference first. An error analysis revealed that this might have something to do with the specificity of our dataset, in that many implicit aspect terms were truly implicit or contained an anaphor used a in a pleonastic-like manner. On the other hand, we also found that even a gold-standard link between an anaphor–antecedent pair can lead to a wrong semantic information link. These findings need to be corroborated on larger and different datasets.

11.3.3 Limitations and future work

Our focus was on the restaurants domain and, in retrospect, it seems that this domain might not have been an excellent choice for exploring the added value of incorporating coreference and semantic role information. In future work, it will thus be very interesting to perform similar experiments on other domains. Currently, however, no other Dutch benchmark review datasets exist.

Regarding the aspect term extraction, we found that the current evaluation metric which takes into account an exact match of the aspect term boundaries is probably too strict. In this respect, it would be interesting to perform a more relaxed evaluation on this specific subtask.

For the aspect category classification experiments, we found that adding lexical-semantic information to the model already helps over relying solely on bag-of-words features. With our Cornetto and DBpedia features we only incorporated a very small amount of such information and in future work it would be very interesting to incorporate more world knowledge.

For the aspect polarity classification we now relied on the entire sentence to assign polarity labels. In future work it will be interesting to use simple window heuristics or dependencies.

11.4 Tipping the scales

To summarize this thesis, we return to our original research question: can coreference resolution and semantic role labeling, i.e. deep semantic processing, lead to better models for automatic readability prediction and aspect-based sentiment analysis? This question calls for a nuanced answer.

Coreference resolution is a hard task in itself, as underlined in the cross-genre robustness experiments. In the extrinsic evaluation on aspect-based sentiment analysis, we found that coreference information only has an added value when gold-standard anaphor-antecedent pairs are used, since this task requires precise predictions. This does not mean, however, that automatic coreference resolution in its current form cannot be of use. For the evaluation on readability prediction, we found that incorporating coreference can indeed be beneficial, mainly because this prediction task is more lenient towards individual errors since information is generalized to the text level.

Contrary to coreference resolution, with semantic role labeling we achieve a moderate performance, and we found that both gold-standard and automatically-obtained semantic roles features boost performance for readability prediction. No positive effect could be observed for the aspect-based sentiment analysis task. Here we expect that the inclusion of more world knowledge into the system will be needed before SRL information can make a difference.

Does deep semantic information help tip the scales on performance? For Dutch readability prediction, we can conclude that it does, when integrated in a state-of-the-art classifier. By using such information for Dutch aspect-based sentiment analysis, we touch the scales but never cause them to tip.

We should stress that these results cannot be generalized to other end-user applications or tasks. Additional research would be required to further assess the impact of deep semantic processing. We hope this dissertation may serve as an inspiration for such future work.

APPENDIX A

Publications

This appendix contains a list of all peer-reviewed journal and conference proceedings publications.

- Articles
  - A1
    - Orphée De Clercq and Véronique Hoste (Under review). All mixed up? Finding the optimal feature set for general readability prediction and its application to English and Dutch. Submitted to Computational Linguistics.
    - Sarah Schulz, Guy De Pauw, Orphée De Clercq, Bart Desmet, Véronique Hoste, Walter Daelemans and Lieve Macken (Under review). Multi-modular text normalization of user-generated content. Submitted to ACM Transactions on Intelligent Systems and Technology.
    - Orphée De Clercq, Véronique Hoste, Bart Desmet, Philip van Oosten, Martine De Cock and Lieve Macken (2014). Using the crowd for readability prediction. Natural Language Engineering, 20 (3). 293-325. Cambridge University Press.
    - Lieve Macken, Orphée De Clercq and Hans Paulussen (2011). Dutch

Parallel Corpus: a balanced copyright-cleared parallel corpus. Meta, 56 (2). 374-390. Les Presses de l'Université de Montréal.

- A3
  - ∗ Maribel Montero Perez, Orphée De Clercq, Piet Desmet, Geert Peeters and Serge Verlinde (2009). Dutch parallel corpus: un nouveau corpus parallèle multilingue disponible en ligne. Romaneske, 34 (4), 2-8.

- Books
  - B2
    - ∗ Orphée De Clercq and Véronique Hoste (2014). Hoe meetbaar is leesbaarheid? In S. Evenepoel, P. Goethals and L. Jooken (eds.), Beschouwingen uit een talenhuis, 147-155. Academia Press, Ghent, Belgium.
    - ∗ Bart Desmet, Orphée De Clercq, Marjan Van de Kauter, Sarah Schulz, Cynthia Van Hee and Véronique Hoste (2014). Taaltechnologie 2.0: sentimentanalyse en normalisatie. In S. Evenepoel, P. Goethals and L. Jooken (eds.), Beschouwingen uit een talenhuis, 157-161. Academia Press, Ghent, Belgium.
    - ∗ Lieve Macken, Orphée De Clercq, Bart Desmet and Véronique Hoste (2014). Dutch Parallel Corpus en SoNaR. In S. Evenepoel, P. Goethals and L. Jooken (eds.), Beschouwingen uit een talenhuis, 163-170. Academia Press, Ghent, Belgium.
- Conference Proceedings
  - P1 (ISI Web of Science)
    - ∗ Orphée De Clercq, Véronique Hoste and Paola Monachesi (2012). Evaluating automatic cross-domain semantic role annotation. Proceedings of the 8th Language Resources and Evaluation Conference (LREC2012), Istanbul, Turkey.
    - ∗ Maaske Treurniet, Orphée De Clercq, Henk van den Heuvel and Nelleke Oostdijk (2012). Collecting a corpus of Dutch SMS. Proceedings of the 8th Language Resources and Evaluation Conference (LREC2012), Istanbul, Turkey.
  - C1
    - ∗ Orphée De Clercq, Marjan Van de Kauter, Els Lefever and Véronique Hoste (2015). LT3: Applying hybrid terminology extraction to Aspect-Based Sentiment Analysis. Proceedings of SemEval2015 Task 12, Denver, USA.
    - ∗ Orphée De Clercq, Michael Schuhmacher, Simone Paolo Ponzetto and Véronique Hoste (2014). Exploiting FrameNet for content-based book recommendation. Proceedings of the CBRecsys workshop at RecSys2014, Foster City, USA. ACM.

- Orphée De Clercq, Sven Hertling, Véronique Hoste, Simone Paolo Ponzetto and Heiko Paulheim (2014). Identifying disputed topics in the news. Proceedings of the LD4KD workshop at ECML/ PKDD2014. CEUR.
- Cynthia Van Hee, Marjan Van de Kauter, Orphée De Clercq, Els Lefever and Véronique Hoste (2014). LT3: Sentiment classification in user-generated content using a rich feature set. Proceedings of SemEval2014 Task 9, Dublin, Ireland.
- Orphée De Clercq, Sarah Schulz, Bart Desmet and Véronique Hoste (2014). Towards shared datasets for normalization research. Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014), Reykjavik, Iceland.
- Orphée De Clercq, Sarah Schulz, Bart Desmet, Els Lefever and Véronique Hoste (2013). Normalisation of Dutch user-generated content. Proceedings of the 9th International Conference on Recent Advances in Natural Language Processing (RANLP2013), Hissar, Bulgaria.
- Orphée De Clercq, Véronique Hoste and Iris Hendrickx (2011). Cross-Domain Dutch coreference resolution. Proceedings of the 8th International Conference on Recent Advances in Natural Language Processing (RANLP 2011), Hissar, Bulgaria.
- Iris Hendrickx, Orphée De Clercq and Véronique Hoste (2011). Analysis and reference resolution of bridge anaphora across different genres. Lecture Notes in Artificial Intelligence, 7099. Springer – Verlag.
- Orphée De Clercq and Maribel Montero Perez (2010). Data collection and IPR in multilingual parallel corpora: Dutch Parallel Corpus. Proceedings of the 7th Language Resources and Evaluation Conference (LREC2010), Valletta, Malta.
- Martin Reynaert, Nelleke Oostdijk, Orphée De Clercq, Henk van den Heuvel and Franciska de Jong (2010). Balancing SoNaR: IPR versus processing issues in a 500-million-word written Dutch reference corpus. Proceedings of the 7th Language Resources and Evaluation Conference (LREC2010), Valletta, Malta.

APPENDIX B

Translations

This appendix contains the translations of the Dutch example snippets that were used for the error analysis of the readability prediction experiments. The numbers correspond to the numbers in the running text.

(19) The Israeli Prime Minister Ariel Sharon prevents the Palestinian president Yasser Arafat from attending the Midnight Mass in Bethlehem. Sharon only wanted to grant Arafat the permission if he arrested the assassins of the Israeli Minister of Tourism Revahan Zeevi. It is the first time that Arafat could not attend the Mass since Bethlehem was turned over to the Palestinian National Authority in 1995. The prohibition provokes criticism. Among others Pope John Paul II, the European Union and the United Nations condemn the measure. A high official of the radical Islamic organisation Jihad announces to end the attacks against Israel. By their own account, Jihad wants to preserve the unity among the Palestinians.

(20) Two teenage girls of 13 and 15 years old jump together from the tenth floor of an apartment building in Brussels. The girls both stayed in a psychiatric institution, the youngest because of a depression. They paid an illicit visit to the father of one of them. When he returned from grocery shopping, he made the grim discovery. According to an inquiry from 1999 by the French-speaking university ULB, Brussels witnesses between 800 and 1,100 suicide

attempts by teenagers between 12 and 19 years old. Suicide has been one of the main causes of death among teenagers in Belgium for years.

(21) The Zimbabwean parliament passes two controversial laws which grant far-reaching power to the government. According to the popular opposition party MCD, president Robert Mugabe wants to eliminate them with those laws in the build-up to the presidential elections in March. Among others, the new laws render criticism against the president illegal. The international community criticises the laws and the policy in Zimbabwe. Especially the rash ousting of white farmers in the land-reform process and the curtailing of the freedom of the press produce negative reactions. Friday January 11 a Zimbabwean delegation will appear before European diplomats in Brussels, who question them with regard to alleged violations of human rights. Europe considers ending the cooperation and financial support of 200 million euros if the situation does not improve. Furthermore, Europe demands the country to allow international observers and press during the elections in March.

(22) Here you see the elements for the answer to your questions. The provided data have not been standardised and do not take the potentially different composition of the age classes and the sex of the populations concerned into account. 1) 8,349 appendectomies were performed for a total sum of 1,653,722 euros in Flanders in 2006. 2) 4,527 appendectomies were performed for a total sum of 902,807 euros in Wallonia in 2006. 3) Based on population data from 2006 which are available at the FPS Economy, SMEs, Self-Employed and Energy there are 137 appendectomies for every 100,000 inhabitants in Flanders, compared to 132 in Wallonia. 4) The costs for the analyses concerning clinical biology for the hospital admissions due to an appendectomy amount to 16.14 euros in Flanders, compared to 23.67 euros in Wallonia.

(23) During the European congress for urology which took place in Istanbul from 16 to 19 March 2005 a session was dedicated to the treatment of stress incontinence through the placement of a suburethral sling. Due to the quality of the sling's texture this surgical technique is used more and more in medicine. The intervention which is carried out first by means of a retropubic approach (TVT) shows excellent results: approximately 90% of the patients regain continence. A new technique was presented which conducts the sling through the foramen obturatorium (TOT, trans obturator tape). That technique is straightforward and does not risk wounding the bladder. This technique does not require performing a cystoscopy first. During this session several groups have compared the 2 techniques. From the various presentations it becomes clear that both surgical techniques can be considered minimally invasive.

(24) Lambermont is often used to refer to the official residence of the Prime

Minister in Belgium (a bit like 10 Downing Street in London). It is located at the corner of Lambermontstraat and Hertogstraat in Brussels, not far from the Royal Palace and the National Palace. This official residence is not to be confused with Wetstraat 16, where the cabinet of the Prime Minister is established. The building is named after baron Auguste Lambermont (1819-1905), who dominated Belgian diplomacy as a hardworking Secretary-General of the Ministry of Foreign Affairs from Brussels during the entire nineteenth century. As a strong advocate of economic liberalism he was one of the driving forces behind the idea of opening the Belgian market for its neighboring countries by means of customs unions. He played among other things a vital role in purchasing the Scheldt toll from the Netherlands (1863) and in the colonial adventures of Leopold II. His talents with regard to negotiations and diplomacy turned out to be essential during various conferences, which established that the independent State of Congo was recognised as the private garden of Leopold II. The building also gave its name to the Lambermont agreement during the state reform of 2001.

(25) The senate committee on Internal Affairs closes the general debate concerning the voting right for immigrants from outside the European Union. Some socialist senators lash out fiercely at VLD president Karel De Gucht, who expressed an absolute 'njet' in the newspapers a few days earlier. The other five coalition partners consider that a diktat. The Flemish liberals especially want to avoid a vote on the proposal. Given the current balance eight out of fifteen committee members would push the green button. As related by senator Jeannine Leduc, VLD betrayed for the first time that the party froze their own proposal with regard to a pentito arrangement so as to prevent the PS from passing the bill which would grant voting rights to migrants. Leduc reminds the Walloon Socialists of that agreement. At the end Philippe Moureaux (PS) is already using conciliatory language. He wants to first resolve the matter within the majority.

(26) The United Nations and the government of Sierra Leone sign an agreement on the establishment of a war tribunal. The court will rule on suspects of atrocities during the civil war which inflicted 50,000 casualties. Last week the end of that war was officially declared. The conflict started in 1991 with actions from the Revolutionary United Front (RUF) under the leadership of Foday Sankoh, later other groups also came into the conflict, foreign ones as well. The RUF became notorious because it mutilated opponents and citizens by chopping their arms off and because it recruited many child soldiers. The new tribunal differs from the other UN tribunals for former Yugoslavia and Rwanda. Those last two were established from within the UN, the establishment of the new tribunal results from a request by the country itself and will consist of UN judges and judges from Sierra Leone.

(27) * You can adjust the tone, the volume and the speed to your own wishes by pushing the buttons up or down. The buttons for tone, volume and speed are situated top center of the surface from left to right, respectively. * To fast forward or backward in a book, press the fast forward or backward button until you have reached the desired position in the book. When you release the button, you will return to the default playback speed automatically.

(28) What benefit has Zyprexa Velotab shown during the studies? Like Zyprexa, Zyprexa Velotab was more effective at improving symptoms than placebo (a dummy treatment). Zyprexa Velotab was as effective as the medicines that it was compared with for the treatment of adults with schizophrenia, the treatment of moderate to severe manic episodes in adults, and the prevention of recurrence in adults with bipolar disorder. What is the risk associated with Zyprexa Velotab? The most common side effects with Zyprexa Velotab (seen in more than 1 patient in 10) are somnolence (sleepiness), weight gain, orthostatic hypotension (sudden drop in blood pressure on standing up) and raised levels of prolactin (a hormone). For the full list of all side effects reported with Zyprexa Velotab, see the package leaflet. Zyprexa Velotab must not be used in people who are hypersensitive (allergic) to olanzapine or any of the other ingredients. It must also not be used in patients at risk of narrow-angle glaucoma (raised pressure inside the eye).

List of Figures

List of Tables

Bibliography

Alderson, J. C.: 1984, *Reading in a Foreign Language: A Reading Problem or a Language Problem*, Longman.

Babko-Malaya, O.: 2005, Propbank annotation guidelines, *Technical report*.

Baccianella, S., Esuli, A. and Sebastiani, F.: 2010, Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining, *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC-2010)*, pp. 2200–2204.

Bagga, A. and Baldwin, B.: 1998, Algorithms for scoring coreference chains, *Proceedings of the 1st International Conference on Language Resources and Evaluation Workshop on Linguistic Coreference (LREC-1998)*, pp. 563–566.

Bailin, A. and Grafstein, A.: 2001, The linguistic assumptions underlying readability formulae: a critique, *Language & Communication* **21**(3), 285–301.

Baker, C., Ellsworth, M. and Erk, K.: 2007, Semeval-2007 task 19: Frame semantic structure extraction, *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pp. 99–104.

Barzilay, R. and Lapata, M.: 2005, Modeling local coherence: an entity-based approach, *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-2005)*, pp. 141–148.

Barzilay, R. and Lapata, M.: 2008, Modeling local coherence: an entity-based approach, *Computational Linguistics* **34**(1), 1–34.

Benjamin, R. G.: 2012, Reconstructing readability: recent developments and recommendations in the analysis of text difficulty, *Educational Psychology Review* **24**(1), 63–88.

Berger, A., Caruana, R., Cohn, D., Freitag, D. and Mittal, V.: 2000, Bridging the lexical chasm: statistical approaches to answer finding, *Proceedings of the International Conference on Research and Development in Information Retrieval (SIGIR-2000)*, pp. 192–199.

Biber, D.: 1988, *Variation Across Speech and Writing*, Cambridge University Press.

Blair-Goldensohn, S., Neylon, T., Hannan, K., Reis, G. A., Mcdonald, R. and Reynar, J.: 2008, Building a sentiment summarizer for local service reviews, *Proceedings of the WWW-2008 workshop on NLP in the Information Explosion Era (NLPIX-2008)*.

Boiy, E. and Moens, M.-F.: 2009, A machine learning approach to sentiment analysis in multilingual web texts, *Information Retrieval* **12**(5), 526–558.

Bouma, G., Daelemans, W., Hendrickx, I., Hoste, V. and Mineur, A.-M.: 2007, The COREA-project, manual for the annotation of coreference in Dutch texts, *Technical report*, University of Groningen.

Briscoe, T. and Carroll, J.: 1997, Automatic Extraction of Subcategorization from Corpora.

Brody, S. and Elhadad, N.: 2010, An unsupervised aspect-sentiment model for online reviews, *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2010)*, pp. 804–812.

Brouwer, R. H. M.: 1963, Onderzoek naar de leesmoeilijkheden van Nederlands proza, *Pedagogische Studiën* **40**, 454–464.

Caro, L. D. and Grella, M.: 2013, Sentiment analysis via dependency parsing, *Computer Standards & Interfaces* **35**(5), 442–453.

Carreras, X. and Màrquez, L.: 2004, Introduction to the CoNLL-2004 Shared Task: Semantic role labeling, *Proceedings of the 8th Conference on Computational Natural Language Learning (CoNLL-2004)*, pp. 89–97.

Carreras, X. and Màrquez, L.: 2005, Introduction to the CoNLL-2005 shared task: Semantic role labeling, *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL-2005)*, pp. 152–164.

Chall, S. and Dale, E.: 1995, *Readability Revisited: The new Dale-Chall Readability Formula*, Brookline Books.

Chang, C.-C. and Lin, C.-J.: 2011, LIBSVM: A library for support vector machines, *ACM Transactions on Intelligent Systems and Technology* **2**(3), 1–27.

Chen, H.-H., Lin, M.-S. and Wei, Y.-C.: 2006, Novel association measures using web search with double checking, *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (C0LING - ACL-2006)*, pp. 1009–1016.

Chinchor, N.: 1998, Overview of MUC-7, *Proceedings of the 7th Message Understanding Conference (MUC-1998).*

Choi, Y., Breck, E. and Cardie, C.: 2006, Joint extraction of entities and relations for opinion recognition, *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP-2006)*, pp. 431–439.

Clark, A. and Toutanova, K.: 2008, *Proceedings of the 12th Conference on Computational Natural Language Learning (CoNLL 2008)*, Coling 2008 Organizing Committee.

Collins-Thompson, K.: 2014, Computational assessment of text readability: a survey of current and future research, *Special Issue of the International Journal of Applied Linguistics* **165**(2), 97–135.

Collins-Thompson, K. and Callan, J.: 2004, A language modeling approach to predicting reading difficulty, *Proceedings of the Human Language Technology Conference and the North American chapter of the Association for Computational Linguistics annual meeting (HLT - NAACL-2004)*, pp. 193–200.

Collins-Thompson, K. and Callan, J.: 2005, Predicting reading difficulty with statistical language models, *Journal of the American Society for Information Science and Technology* **56**, 1448–1462.

Connolly, D., Burger, J. and Day, D.: 1994, A Machine learning approach to anaphoric reference, *Proceedings of the International Conference on New Methods in Language Processing (NeMLaP)*, pp. 255–261.

Cristianini, N. and Shawe-Taylor, J.: 2000, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press.

Crossley, S., Greenfield, J. and McNamara, D.: 2008, Assessing text readability using cognitively based indices, *TESOL Quarterly* **43**(3), 475–493.

Dabrowski, M., Acton, T., Jarzebowski, P. and O'Riain, S.: 2010, Improving customer decisions using product reviews - CROM - Car Review Opinion Miner, *Proceedings of the 6th International Conference on Web Information Systems and Technologies (WEBIST-2010)*, pp. 354–357.

Daelemans, W. and van den Bosch, A.: 2005, *Memory-based Language Processing*, Cambridge University Press.

Daelemans, W., Zavrel, J., van den Bosch, A. and van der Sloot, K.: 2003, MBT: Memory Based Tagger, version 2.0, Reference Guide, *Technical Report ILK Research Group Technical Report Series no. 03-13*, ILK Research Group, Tilburg University.

Daelemans, W., Zavrel, J., van der Sloot, K. and van den Bosch, A.: 2009, TiMBL: Tilburg Memory Based Learner, version 6.2, Reference Guide, *Technical Report 09-01*, ILK Research Group, Tilburg University.

Daelemans, W., Zavrel, J., Van der Sloot, K. and van den Bosch, A.: 2010, TiMBL: Tilburg Memory Based Learner, version 6.3, Reference Guide, *Technical Report 10-01*, ILK Research Group, Tilburg University.

Daumé III, H., Deoskar, T., McClosky, D., Plank, B. and Tiedemann, J. (eds): 2010, *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing (DANLP-2010)*, Association for Computational Linguistics.

Davison, A. and Kantor, R.: 1982, On the failure of readability formulas to define readable texts: a case study from adaptations, *Reading Research Quarterly* **17**(2), 187–209.

De Clercq, O., Desmet, B., Schulz, S., Lefever, E. and Hoste, V.: 2013, Normalization of Dutch user-generated content, *Proceedings of Recent Advances in Natural Language Processing*, INCOMA, pp. 179–188.

De Clercq, O., Hendrickx, I. and Hoste, V.: 2011, Cross-domain Dutch coreference resolution, *Proceedings of the 8th International Conference on Recent Advances in Natural Language Processing (RANLP-2011)*, pp. 186–193.

De Clercq, O. and Hoste, V.: under review, All mixed up? Finding the optimal feature set for general readability prediction and its application to English and Dutch, *Computational Linguistics* .

De Clercq, O., Hoste, V., Desmet, B., van Oosten, P., De Cock, M. and Macken, L.: 2014, Using the crowd for readability prediction, *Natural Language Engineering, Cambridge Journals Online* **20**(3), 293–325.

De Clercq, O., Monachesi, P. and Hoste, V.: 2012, Evaluating automatic cross-domain semantic role annotation, *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012)*, pp. 88–93.

De Clercq, O., Van de Kauter, M., Lefever, E. and Hoste, V.: 2015, LT3: Applying hybrid terminology extraction to aspect-based sentiment analysis, *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval-2015)*, pp. 719–724.

de Salvo Braz, R., Girju, R., Punyakanok, V., Roth, D. and Sammons, M.: 2005, An inference model for semantic entailment in natural language, *Proceedings of the 20th National Conference on Artificial Intelligence (AAAI-2005)*, pp. 1043–1049.

De Smedt, T. and Daelemans, W.: 2012, Vreselijk mooi! Terribly beautiful: a subjectivity lexicon for Dutch adjectives, *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012)*, pp. 3568–3572.

Denis, P. and Baldridge, J.: 2008, Specialized models and ranking for coreference resolution, *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP-2008)*, pp. 660–669.

Desmet, B.: 2014, *Finding the online cry for help: automatic text classification for suicide prevention*, Ghent University.

Desmet, B. and Hoste, V.: 2013, Fine-grained Dutch named entity recognition, *Language Resources and Evaluation* pp. 307–343.

Desmet, B., Hoste, V., Verstraeten, D. and Verhasselt, J.: 2013, Gallop Documentation, *Technical Report LT3 13-03*, Ghent University.

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S. and Weischedel, R.: 2004, The Automatic Content Extraction (ACE) Program Tasks, Data, and Evaluation, *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004)*, pp. 837–840.

Douma, W.: 1960, De leesbaarheid van landbouwbladen: een onderzoek naar en een toepassing van leesbaarheidsformules, *Bulletin* **17**.

Dowty, D.: 1991, Thematic proto-roles and argument selection, *Language* **67**, 547–619.

DuBay, W. H.: 2004, *The Principles of Readability*, Impact Information.

DuBay, W. H. (ed.): 2007, *Unlocking Language: the Classic Readability Studies*, BookSurge.

Feldman, R.: 2013, Techniques and applications for sentiment analysis, *Communications of the ACM* **56**(4), 82–89.

Fellbaum, C.: 1998, *WordNet: an Electronic Lexical Database*, MIT Press.

Feng, L.: 2010, *Automatic Readability Assessment*, Phd, The City University of New York.

Feng, L., Elhadad, N. and Huenerfauth, M.: 2009, Cognitively motivated features for readability assessment, *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2009)*, pp. 229–237.

Feng, L., Jansche, M., Huenerfauth, M. and Elhadad, N.: 2010, A comparison of features for automatic readability assessment, *Proceedings of the 23rd International Conference on Computational Linguistics Poster Volume (COLING-2010)*, pp. 276–284.

Ferrucci, D. A., Brown, E. W., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A., Lally, A., Murdock, J. W., Nyberg, E., Prager, J. M., Schlaefer, N. and Welty, C. A.: 2010, Building watson: An overview of the DeepQA project, *AI Magazine* **31**(3), 59–79.

Fillmore, C. J., Johnson, C. R. and Petruck, M. R. L.: 2003, Background to FrameNet, *International Journal of Lexicography* **16**(3), 235–250.

Fillmore, C. J., Ruppenhofer, J. and Baker, C. F.: 2004, FrameNet and representing the link between semantic and syntactic relations, *in* C.-r. Huang and W. Lenders (eds), *Computational Linguistics and Beyond*, Language and Linguistics Monographs Series B, Institute of Linguistics, Academia Sinica, pp. 19–62.

Flesch, R.: 1948, A new readability yardstick, *Journal of Applied Psychology* **32**(3), 221–233.

Foster, J., Cetinoglu, O., Wagner, J., Roux, J. L., Hogan, S., Nivre, J., Hogan, D. and van Genabith, J.: 2011, #hard-toparse: POS tagging and parsing the twitterverse, *Proceedings of the 25th National Conference on Artificial Intelligence Workshop on Analyzing Microtext (AAAI-2011)*, pp. 20–25.

François, T.: 2009, Combining a statistical language model with logistic regression to predict the lexical and syntactic difficulty of texts for FFL, *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop (EACL-2009)*, pp. 19–27.

François, T.: 2011, *Les apports du traitement automatique du language à la lisibilite du français langue etrangère*, PhD thesis, Louvain-La-Neuve.

François, T. and Miltsakaki, E.: 2012, Do NLP and machine learning improve traditional readability formulas?, *Proceedings of the 1st Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR2012)*, pp. 49–57.

Frantzi, K. and Ananiadou, S.: 1999, The C-value / NC-value domain independent method for multi-word term extraction, *Journal of Natural Language Processing* **6**(3), 145–179.

Galley, M. and Mckeown, K.: 2003, Improving word sense disambiguation in lexical chaining, *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-2003)*, pp. 1486–1488.

Gamon, M., Aue, A., Corston-Oliver, S. and Ringger, E.: 2005, Pulse: Mining customer opinions from free text, *Proceedings of the 6th International Conference on Advances in Intelligent Data Analysis (IDA-2005)*, pp. 121–132.

Gantz, J. and Reinsel, D.: 2011, Extracting value from chaos, *Technical report*, Framingham, MA: International Data Corporation (IDC).

Ganu, G., Elhadad, N. and Marian, A.: 2009, Beyond the stars: improving rating predictions using review text content, *Proceedings of the 12th International Workshop on the Web and Databases (WebDB-2009)*, pp. 1–6.

Gao, Q. and Vogel, S.: 2011, Corpus expansion for statistical machine translation with semantic role label substitution rules, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Short Papers (ACL-2011)*, pp. 294–298.

Gildea, D. and Jurafsky, D.: 2002, Automatic labeling of semantic roles, *Computational Linguistics* **28**, 245–288.

Goldberg, D.: 1989, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison Wesley.

Goldberg, D. and Deb, K.: 1991, A comparative analysis of selection schemes used in genetic algorithms, *Foundations of Genetic Algorithms*, Morgan Kaufmann Publishers, pp. 69–93.

Graesser, A. C., McNamara, D. S., Louwerse, M. M. and Cai, Z.: 2004, Cohmetrix: analysis of text on cohesion and language, *Behavior Research Methods, Instruments and Computers* **36**, 193–202.

Gunning, R.: 1952, *The Technique of Clear Writing*, McGraw-Hill, New York.

Haghighi, A. and Klein, D.: 2009, Simple coreference resolution with rich syntactic and semantic features, *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP-2009)*, pp. 1152–1161.

Hajič, J.: 2009, *Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL-2009): Shared Task*, Association for Computational Linguistics.

Hajič, J., Panevová, J., Hajičová, E., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M., Žabokrtský, Z., Ševčíková Razímová, M. and Urešová, Z.: 2006, Prague dependency treebank 2.0 (PDT 2.0).

Halliday, M. and Hasan, R.: 1976, *Cohesion in English*, Longman Group Ltd.

Hatzivassiloglou, V. and McKeown, K. R.: 1997, Predicting the semantic orientation of adjectives, *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the Association for Computational Linguistics (ACL - EACL-1997)*, pp. 174–181.

Heilman, M., Collins-Thompson, K. and Eskenazi, M.: 2008, An analysis of statistical models and features for reading difficulty prediction, *Proceedings of the 3rd ACL Workshop on Innovative Use of NLP for Building Educational Applications (EANL-2008)*, pp. 71–79.

Heilman, M. J., Collins-Thompson, K., Callan, J. and Eskenazi, M.: 2007, Combining lexical and grammatical features to improve readability measures for first and second language texts, *Proceedings of the Human Language Technology Conference and the North American chapter of the Association for Computational Linguistics annual meeting (HLT - NAACL 2007)*, pp. 460–467.

Hendrickx, I., Bouma, G., Coppens, F., Daelemans, W., Hoste, V., Kloosterman, G., Mineur, A.-M., Van Der Vloet, J. and Verschelde, J.-L.: 2008, A coreference corpus and resolution system for Dutch, *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2008)*, pp. 144–149.

Hendrickx, I., Bouma, G., Daelemans, W. and Hoste, V.: 2013, COREA: Coreference Resolution for Extracting Answers for Dutch, *in* P. Spyns and J. Odijk (eds), *Essential Speech and Language Technology for Dutch*, Theory and Applications of Natural Language Processing, Springer, pp. 115–128.

Hendrickx, I. and Hoste, V.: 2009, Coreference resolution on blogs and commented news, *Anaphora Processing and Applications, Lecture Notes in Artificial Intelligence*, Vol. 5847, pp. 43–53.

Hoste, V.: 2005, *Optimization Issues in Machine Learning of Coreference Resolution*, PhD thesis, Antwerp University.

Hoste, V. and De Pauw, G.: 2006, KNACK-2002: A richly annotated corpus of Dutch written text, *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*, pp. 1432–1437.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L. and Weischedel, R.: 2006, OntoNotes: The 90% solution, *Proceedings of the Human Language Technology Conference and the North American chapter of the Association for Computational Linguistics annual meeting: Short Papers (HLT - NAACL 2006)*, pp. 57–60.

Hovy, E., Navigli, R. and Ponzetto, S. P.: 2013, Collaboratively built semi-structured content and Artificial Intelligence: The story so far, *Artificial Intelligence* **194**, 2–27.

Hsu, C.-W., Chang, C.-C. and Chih-Jen, L.: 2003, A practical guide to support vector classification, *Technical Report National Taiwan University*, Department of Computer Science, National Taiwan University.

Hu, M. and Liu, B.: 2004, Mining and summarizing customer reviews, *Proceedings of the 10th International Conference on Knowledge Discovery and Data Mining (KDD-2004)*, pp. 168–177.

Iida, R., Inui, K., Takamura, H. and Matsumoto, Y.: 2003, Incorporating contextual cues in trainable models for coreference resolution, *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics Workshop on The Computational Treatment of Anaphora (EACL-2003)*, pp. 23–30.

Jackendoff, R.: 1990, *Semantic Structures*, MIT Press.

Jakob, N. and Gurevych, I.: 2010, Using anaphora resolution to improve opinion target identification in movie reviews, *Proceedings of the 8th Annual Meeting of the Association for Computational Linguistics (ACL-2010)*, pp. 263–268.

Jiang, L., Yu, M., Zhou, M., Liu, X. and Zhao, T.: 2011, Target-dependent twitter sentiment classification, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-2011)*, pp. 151–160.

Jijkoun, V. and Hofmann, K.: 2009, Generating a non-English subjectivity lexicon: Relations that matter, *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2009)*, pp. 398–405.

Johansson, P. and Nugues, P.: 2008, The effect of syntactic representation on semantic role labeling, *Proceedings of the 22nd International Conference on Computational Linguistics (CLING-2008)*, pp. 393–400.

Johansson, R. and Moschitti, A.: 2013, Relational features in fine-grained opinion analysis, *Computational Linguistics* **39**(3), 473–509.

Jurafsky, D. and Martin, J. H.: 2008, *Speech and Language Processing*, Prentice Hall.

Justeson, K. and Katz, S.: 1995, Technical terminology: some linguistic properties and an algorithm for identification in text, *Natural Language Engineering* **1**(1), 9–27.

Kageura, K. and Umino, B.: 1996, Methods of automatic term recognition. A review, *Terminology* **3**(2), 259–289.

Karttunen, L.: 1976, Discourse referents, *Syntax and Semantics* **7**.

Kate, R. J., Luo, X., Patwardhan, S., Franz, M., Florian, R., Mooney, R. J., Roukos, S. and Welty, C.: 2010, Learning to predict readability using diverse linguistic features, *Proceedings of the 23rd International Conference on Computational Linguistics (COLING-2010)*, pp. 546–554.

Kessler, J. S. and Nicolov, N.: 2009, Targeting sentiment expressions through supervised ranking of linguistic configurations, *The 3rd Association for the Advancement of Artificial Intelligence Conference on Weblogs and Social Media (ICWSM-2009)*, pp. 90–97.

Kim, S.-M. and Hovy, E.: 2006, Extracting opinions, opinion holders, and topics expressed in online news media text, *Proceedings of the Workshop on Sentiment and Subjectivity in Text (SST-2006)*, pp. 1–8.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L. and Chissom., B. S.: 1975, Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel, *Research branch report RBR-8-75*, Naval Technical Training Command Millington Tenn Research Branch, Springfield, Virginia.

Klare, G. R.: 1976, A second look at the validity of readability formulas, *Journal of Literacy Research* **8**, 129–152.

Kobayashi, N., Inui, K. and Matsumoto, Y.: 2007, Extracting aspect-evaluation and aspect-of relations in opinion mining, *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP - CoNLL-2007*, pp. 1065–1074.

Koller, A. and Pinkal, M.: 2012, Semantic research in computational linguistics, *in* C. Maienborn, K. von Heusinger and P. Portner (eds), *Semantics: An International Handbook of Natural Language Meaning*, Mouton de Gruyter, pp. 2825–2858.

Kraf, R. and Pander Maat, H.: 2009, Leesbaarheidsonderzoek: oude problemen, nieuwe kansen, *Tijdschrift voor Taalbeheersing* **31**(2), 97–123.

Lee, D. Y. W.: 2001, Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the bnc jungle, *Language Learning & Technology* **5**(3), 37–72.

Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M. and Jurafsky, D.: 2013, Deterministic coreference resolution based on entity-centric, precision-ranked rules, *Computational Linguistics* **39**(4), 885–916.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S. and Bizer, C.: 2013, DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia, *Semantic Web Journal* **6**, 167–195.

Leroy, G. and Endicott, J.: 2011, Term familiarity to indicate perceived and actual difficulty of text in medical digital libraries, *International Conference on Asia-Pacific Digital Libraries (ICADL-2011)*, pp. 307–310.

Leroy, G., Helmreich, S., Cowie, J. R., Miller, T. and Zheng, W.: 2008, Evaluating online health information: beyond readability formulas, *Proceedings of the American Medical Informatics Association Annual Symposium (AMIA-2008)*, pp. 394–398.

Li, B., Zhou, L., Feng, S. and Wong, K.-F.: 2010, A unified graph model for sentence-based opinion retrieval, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-2010)*, pp. 1367–1375.

Lin, D. and Wu, X.: 2009, Phrase clustering for discriminative learning, *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-2009)*, pp. 1030–1038.

Litkowski, K.: 2004, Senseval-3 task: Automatic labeling of semantic roles, *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3)*, pp. 9–12.

Liu, B.: 2012, Sentiment analysis and opinion mining, *Synthesis Lectures on Human Language Technologies* **5**(1), 1–167.

Liu, B., Hu, M. and Cheng, J.: 2005, Opinion observer: Analyzing and comparing opinions on the web, *Proceedings of the 14th International Conference on World Wide Web (WWW-2005)*, pp. 342–351.

Liu, D. and Gildea, D.: 2010, Semantic role features for machine translation, *Proceedings of the 23rd International Conference on Computational Linguistics (COLING-2010)*, pp. 716–724.

Liu, Y. and Lin, S.: 2005, Log-linear models for word alignment, *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-2005)*, pp. 459–466.

Luo, X., Ittycheriah, A., Jing, H., Kambhatla, N. and Roukos, S.: 2004, A mention-synchronous coreference resolution algorithm based on the bell tree, *in* S. Barcelona (ed.), *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-2004)*, pp. 136–143.

Luo, X. and Zitouni, I.: 2005, Multi-lingual coreference resolution with syntactic features, *Proceedings of the Human Language Technology Conference and the 2005 Conference on Empirical Methods in Natural Language Processing (HLT - EMNLP-2005)*, pp. 660–667.

Macken, L., Lefever, E. and Hoste, V.: 2013, TExSIS: Bilingual Terminology Extraction from Parallel Corpora Using Chunk-based Alignment, *Terminology* **19**(1), 1–30.

Maks, I., Martin, W. and de Meerseman, H.: 1999, RBN Manual, *Technical report*, Vrije Universiteit Amsterdam.

Manning, C. D., Raghavan, P. and Schütze, H.: 2008, *Introduction to Information Retrieval*, Cambridge University Press. online version: http://nlp.stanford.edu/IR-book/html/htmledition/mybook.html.

Marcus, M. P., Santorini, B. and Marcinkiewicz, M. A.: 1993, Building a large annotated corpus of English: the Penn Treebank, *Computational Linguistics* **19**, 313–330.

Màrquez, L., Carreras, X., Litkowski, K. C. and Stevenson, S.: 2008, Semantic role labeling: An introduction to the special issue, *Computational Linguistics* **34**(2).

Martin, W.: 2005, Referentiebestand Nederlands, *Technical report*, Vrije Universiteit Amsterdam.

Martin, W. and Ploeger, J.: 1999, Tweetalige woordenboeken voor het Nederlands: het beleid van de commissie lexicografische vertaalvoorzieningen, *Neerlandica Extra Muros* **37**, 22–32.

McNamara, D. S., Louwerse, M. M., McCarthy, P. M. and Graesser, A. C.: 2010, Coh-metrix: Capturing linguistic features of cohesion, *Discourse Processes* **47**(4), 292–330.

Melli, G., Wang, Y., Liu, Y., Kashani, M. M., Shi, Z., Gu, B., Sarkar, A. and Popowich, F.: 2005, Description of SQUASH, the SFU Question Answering Summary Handler for the duc-2005 summarization task, *Proceedings of the Document Understanding Conference (DUC-2005)*.

Mendes, P. N., Jakob, M., García-Silva, A. and Bizer, C.: 2011, DBpedia Spotlight: Shedding light on the web of documents, *Proceedings of the 7th International Conference on Semantic Systems (I-Semantics-2011)*, pp. 1–8.

Merlo, P. and Stevenson, S.: 2001, Automatic verb classification based on statistical distributions of argument structure, *Computational Linguistics* **27**(3), 373–408.

Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B. and Grishman, R.: 2004, Annotating noun argument structure for nombank, *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004)*, pp. 803–806.

Mikolov, T., Chen, K., Corrado, G. and Dean, J.: 2013, Efficient estimation of word representations in vector space, *CoRR* .

Mitchell, M.: 1996, *An Introduction to Genetic Algorithms*, MIT Press.

Mitkov, R.: 2002, *Anaphora Resolution*, Longman.

Mitkov, R., Evans, R., Orasan, C., Dornescu, I. and Rios, M.: 2012, Coreference resolution: To what extent does it help NLP applications?, *Proceedings of the 15th International Conference on Text, Speech and Dialogue (TSD-2012)*, pp. 16–27.

Moens, M.-F., Li, J. and Chua, T.-S. (eds): 2014, *Mining user generated content*, Chapman and Hall/CRC.

Monachesi, P., Stevens, G. and Trapman, J.: 2007, Adding semantic role annotation to a corpus of written Dutch, *Proceedings of the Linguistic Annotation Workshop*, pp. 77–84.

Mooney, R.: 1996, Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning, *Proceedings of the 1996 Conference on Empirical Methods in Natural Language Processing (EMNLP-1996)*, pp. 82–91.

Morton, T. S.: 2000, Coreference for NLP applications, *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL-2000)*, pp. 173–180.

Moschitti, A. and Basili, R.: 2004, Complex linguistic features for text classification: A comprehensive study, *Proceedings of the 26th European Conference on Information Retrieval (ECIR-2004)*, pp. 181–196.

MUC-6: 1995, Coreference task definition. Version 2.3., *Proceedings of the 6th Message Understanding Conference (MUC-6)*, pp. 335–344.

Narayanan, S. and Harabagiu, S.: 2004, Question answering based on semantic structures, *Proceedings of the 20th International Conference on Computational Linguistics(COLING-2004)*, pp. 59–65.

Nenkova, A., Chae, J., Louis, A. and Pitler, E.: 2010, Structural features for predicting the linguistic quality of text: Applications to machine translation, automatic summarization and human-authored text, *Empirical Methods in NLG, Lecture Notes in Artificial Intelligence* **5790**, 222–241.

Ng, V.: 2010, Supervised noun phrase coreference research: The first fifteen years, *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL-2010)*, pp. 1396–1411.

Ng, V. and Cardie, C.: 2002, Improving machine learning approaches to coreference resolution, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*, pp. 104–111.

Nicolov, N., Salvetti, F. and Ivanova, S.: 2008, Sentiment analysis: Does coreference matter?, *Proceedings of the Symposium on Affective Language in Human and Machine*, pp. 37–40.

Noreen, E.: 1989, *Computer Intensive Methods for Testing Hypothesis: An Introduction*, John Wiley & Sons, New York.

OECD: 2013, OECD SKills Outlook 2013, *Technical report*, OECD Publishing.

O'Hare, N., Davy, M., Bermingham, A., Ferguson, P., Sheridan, P., Gurrin, C. and Smeaton, A. F.: 2009, Topic-dependent sentiment analysis of financial blogs, *Proceedings of the 1st International Conference on Information and Knowledge Management Workshop on Topic-sentiment Analysis for Mass Opinion (TSA-2009)*, pp. 9–16.

Oostdijk, N., Reynaert, M., Hoste, V. and Schuurman, I.: 2013, The construction of a 500-million-word reference corpus of contemporary written Dutch, *Essential Speech and Language Technology for Dutch, Theory and Applications of Natural Language Processing*, Springer, pp. 219–247.

Palmer, M., Gildea, D. and Kingsbury, P.: 2005, The proposition bank: A corpus annotated with semantic roles, *Computational Linguistics Journal* **31**(1), 71–106.

Pang, B. and Lee, L.: 2008, Opinion mining and sentiment analysis, *Found. Trends Inf. Retr.* **2**(1-2).

Pang, B., Lee, L. and Vaithyanathan, S.: 2002, Thumbs up?: Sentiment classification using machine learning techniques, *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*, pp. 79–86.

Paulheim, H. and Fürnkranz, J.: 2012, Unsupervised Generation of Data Mining Features from Linked Open Data, *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics (WIMS-2012)*, p. 31.

Pavlopoulos, I.: 2014, *Aspect Based Sentiment Analysis*, Phd, Department of Informatics, Athens University of Economics and Business.

Pavlopoulos, J. and Androutsopoulos, I.: 2014, Aspect term extraction for sentiment analysis: New datasets, new evaluation measures and an improved unsupervised method, *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM-2014)*, pp. 44–52.

Petersen, S. and Ostendorf, M.: 2009, A machine learning approach to reading level assessment, *Computer Speech & Language* **23**(1), 89–106.

Pitler, E. and Nenkova, A.: 2008, Revisiting readability: A unified framework for predicting text quality, *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP-2008)*, ACL, pp. 186–195.

Poesio, M., Ponzetto, S. and Versley, Y.: 2010, *Computational models of anaphora resolution: A survey*, http://clic.cimec.unitn.it/massimo/Publications/lilt.pdf.

Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S. and Androutsopoulos, I.: 2015, Semeval-2015 task 12: Aspect based sentiment analysis, *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval-2015)*, pp. 486–495.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I. and Manandhar, S.: 2014, Semeval-2014 task 4: Aspect based sentiment analysis, *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, pp. 27–35.

Popescu, A.-M. and Etzioni, O.: 2005, Extracting product features and opinions from reviews, *Proceedings of the 2005 Conference on Human Language Technology and Empirical Methods in Natural Language Processing (EMNLP-2005)*, pp. 339–346.

Pradhan, S., Moschitti, A., Xue, N., Uryupina, O. and Zhang, Y.: 2012, CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes, *Proceedings of the Sixteenth Conference on Computational Natural Language Learning (CoNLL-2012)*, pp. 1–40.

Pradhan, S., Ward, W. and Martin, J.: 2008, Towards robust semantic role labeling, *Computational Linguistics* **34**, 289–310.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. and Webber, B.: 2008, The penn discourse treebank 2.0, *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC-2008)*.

Rahman, A. and Ng, V.: 2009, Supervised models for coreference resolution, *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP-2009)*, pp. 968–977.

Rand, W. M.: 1971, Objective criteria for the evaluation of clustering methods, *Journal of the American Statistical Association* **66**(336), 846–850.

Rayson, P. and Garside, R.: 2000, Comparing corpora using frequency profiling, *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics Workshop on Comparing Corpora, (ACL-2000)*, pp. 1–6.

Recasens, M. and Hovy, E.: 2011, BLANC: Implementing the rand index for coreference evaluation, *Natural Language Engineering,* pp. 485–510.

Recasens, M., Márquez, L., Sapena, E., Martí, M. A., Tauleé, M., Hoste, V., Poesio, M. and Versley, Y.: 2010, SemEval-2010 Task 1: Coreference resolution in multiple languages, *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010)*, pp. 1–8.

Recasens, M. and Martí, M. A.: 2010, AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan, *Language Resources and Evaluation* **44**(4), 315–345.

Reeve, L. and Han, H.: 2005, Survey of semantic annotation platforms, *Proceedings of the 2005 Asoociation for Computing Machinery Symposium on Applied Computing (SAC-2005)*, pp. 1634–1638.

Ritter, A., Clark, S., Mausam and Etzioni, O.: 2011, Named entity recognition in tweets: An experimental study, *Proceedings of the 2011 Conference on Empirical Methods for Natural Language Processing EMNLP-2011*, pp. 1524–1534.

Ruppenhofer, J., Somasundaran, S. and Wiebe, J.: 2008, Finding the sources and targets of subjective expressions, *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC-2008)*, pp. 2781–2788.

Sabou, M., Bontcheva, K., Derczynski, L. and Scharl, A.: 2014, Corpus annotation through crowdsourcing: Towards best practice guidelines, *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC-2014)*, pp. 859–866.

Sabou, M., Bontcheva, K. and Scharl, A.: 2012, Crowdsourcing research opportunities: Lessons from natural language processing, *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies (i-Know-2012)*.

Saias, J.: 2015, Sentiue: Target and aspect based sentiment analysis in SemEval-2015 Task 12, *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval-2015)*, pp. 767–771.

Salton, G.: 1989, *Automatic text processing: The transformation, analysis and retrieval of information by computer*, Addison Wesley.

Sammons, M., Vydiswaran, V., Vieira, T., Johri, N., Chang, M., Goldwasser, D., Srikumar, V., Kundu, G., Tu, Y., Small, K., Rule, J., Do, Q. and Roth, D.: 2009, Relation alignment for textual entailment recognition, *Proceedings of the 2009 Text Analysis Conference (TAC-2009)*.

Samuels, S. J., Zakaluk, B. L. and Association, I. R.: 1988, *Readability: Its past, present, and future*, Newark, Del. : International Reading Association.

San Vicente, I. n., Saralegi, X. and Agerri, R.: 2015, Elixa: A modular and flexible ABSA platform, *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval-2015)*, pp. 748–752.

Schulte im Walde, S.: 2006, Experiments on the automatic induction of German semantic verb classes, *Computational Linguistics* **32**(2), 159–194.

Schuurman, I., Hoste, V. and Monachesi, P.: 2010, Interacting semantic layers of annotation in SoNaR, a reference corpus of contemporary written Dutch, *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC-2010)*, pp. 2471–2477.

Schwarm, S. E. and Ostendorf, M.: 2005, Reading level assessment using support vector machines and statistical language models, *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-2005)*, pp. 523–530.

Shadbolt, N., Berners-Lee, T. and Hall, W.: 2006, The semantic web revisited, *Institute of Electrical and Electronics Engineers Intelligent Systems* **21**(3), 96–101.

Shanahan, J. G., Qu, Y. and Wiebe, J. (eds): 2006, *Computing Attitude and Affect in Text: Theory and Applications*, number 20 in *the Information Retrieval Series*, Springer.

Shen, D. and Lapata, M.: 2007, Using semantic roles to improve question answering, *Proceedings of the 2007 Conference on Empirical Methods in Natural Language Processing and on Computational Natural Language Learning (EMNLP-CoNLL-2007)*, pp. 12–21.

Si, L. and Callan, J.: 2001, A statistical model for scientific readability, *Proceedings of the 10th International Conference on Information Knowledge Management (ICKM-2001)*, pp. 574–576.

Snow, R., O'Connor, B., Jurafsky, D. and Ng, A. Y.: 2008, Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks, *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP-2008)*, pp. 254–263.

Soon, W. M., Ng, H. T. and Lim, D. C. Y.: 2001, A Machine Learning Approach to Coreference Resolution of Noun Phrases, *Computational Linguistics* **27**(4), 521–544.

Sprenger, T. O., Tumasjan, A., Sandner, P. G. and Welpe, I. M.: 2014, Tweets and trades: The information content of stock microblogs, *European Financial Management* **20**(5), 926–957.

Spyns, P. and Odijk, J.: 2013, *Essential Speech and Language Technology for Dutch, Theory and Applications of Natural Language Processing*, Springer Berlin Heidelberg.

Staphorsius, G.: 1994, *Leesbaarheid en leesvaardigheid. De ontwikkeling van een domeingericht meetinstrument*, Cito, Arnhem.

Staphorsius, G. and Krom, R. S.: 1985, *Cito leesbaarheidsindex voor het basisonderwijs: Verslag van een leesbaarheidsonderzoek*, number 36 in *Specialistisch bulletin*, Cito Arnhem.

Steinberger, J., Poesio, M., Kabadjov, M. and Jezek., K.: 2007, Two uses of anaphora resolution in summarization, *Information Processing and Management. Special issue on Summarization* **43**, 1663–1680.

Stenetorp, P., Pyysalo, S., Topic, G., Ohta, T., Ananiadou, S. and Tsujii, J.: 2012, BRAT: a web-based tool for NLP-assisted text annotation, *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2012)*, pp. 102–107.

Stevens, G., Monachesi, P. and van den Bosch, A.: 2007, A pilot study for automatic semantic role labeling in a Dutch corpus, *Selected papers from the seventeenth CLIN meeting*, LOT Occasional Series 7.

Stolcke, A.: 2002, SRILM - an extensible language modeling toolkit, *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP-2002)*, pp. 901–904.

Stoyanov, V. and Cardie, C.: 2006, Partially supervised coreference resolution for opinion summarization through structured rule learning, *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP-2006)*, pp. 336–344.

Stoyanov, V. and Cardie, C.: 2008, Topic identification for fine-grained opinion analysis, *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-2008)*, pp. 817–824.

Strube, M.: 2009, Anaphernresolution, *in* K.-U. Carstensen, C. Ebert, C. Endriss, S. Jekat, R. Klabunde and H. Langer (eds), *Computerlinguistik und Sprachtechnologie. Eine Einfuhrung*, Springer Germany, pp. 399–409.

Surdeanu, M., Harabagiu, S., Williams, J. and Aarseth, P.: 2003, Using predicate-argument structures for information extraction, *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-2003)*, pp. 8–15.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. and Stede, M.: 2011, Lexicon-based methods for sentiment analysis, *Computational Linguistics* **37**(2), 267–307.

Tanaka-Ishii, K., Tezuka, S. and Terada, H.: 2010, Sorting texts by readability, *Computational Linguistics* **36**(2), 203–227.

Telljohann, H., Hinrichs, E. and Kübler, S.: 2004, The TüBa-D/Z treebank: Annotating German with a context-free backbone, *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC-2004)*, pp. 2229–2235.

Thet, T. T., Na, J.-C. and Khoo, C. S.: 2010, Aspect-based sentiment analysis of movie reviews on discussion boards, *Journal of Information Science* **36**(6), 823–848.

Titov, I. and McDonald, R.: 2008, A joint model of text and aspect ratings for sentiment summarization, *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL-2008)*, pp. 308–316.

Tjong Kim Sang, E.: 2002, Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition, *Proceedings of the 6th Conference on Natural Language Learning (COLING-2002)*, pp. 155–158.

Todirascu, A., François, T., Gala, N., Fairon, C., Ligozat, A.-L. and Bernhard, D.: 2013, Coherence and cohesion for the assessment of text readability, *Proceedings of 10th International Workshop on Natural Language Processing and Cognitive Science (NLPCS-2013)*, pp. 11–19.

Toh, Z. and Su, J.: 2015, NLANGP: Supervised machine learning system for aspect category classification and opinion target extraction, *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval-2015)*, pp. 496–501.

Toh, Z. and Wang, W.: 2014, DLIREC: Aspect term extraction and term polarity classification system, *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, pp. 235–240.

Trapman, J. and Monachesi, P.: 2006, Manual for semantic annotation in D-Coi, *Technical report*, Utrecht University, Uil-OTS.

van Boom, W.: 2014, Begrijpelijke hypotheekvoorwaarden en consumentengedrag, *in* T. B. en A.A. van Velten (ed.), *Perspectieven voor vastgoedfinanciering (Congresbundel Stichting Fundatie Bachiene)*, Stichting Fundatie Bachiene, pp. 45–80.

Van de Cruys, T.: 2005, Semantic clustering in Dutch, *Proceedings of the Sixteenth Computational Linguistics in the Netherlands (CLIN)*, pp. 17–32.

Van de Kauter, M., Coorman, G., Lefever, E., Desmet, B., Macken, L. and Hoste, V.: 2013, LeTs Preprocess: The multilingual LT3 linguistic preprocessing toolkit, *Computational Linguistics in the Netherlands Journal* **3**, 103–120.

van den Bosch, A., Busser, B., Daelemans, W. and Canisius, S.: 2007, An efficient memory-based morphosyntactic tagger and parser for Dutch, *Proceedings of the Seventeenth Computational Linguistics in the Netherlands (CLIN)*, pp. 191–206.

Van Eynde, F.: 2005, Part of speech tagging en lemmatisering van het D-Coi Corpus, *Technical report*, Centrum voor Computerlinguïstiek, KU Leuven.

Van Hee, C., Van de Kauter, M., De Clercq, O., Lefever, E. and Hoste, V.: 2014, LT3: Sentiment classification in user-generated content using a rich feature set, *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval-2014)*, pp. 406–410.

van Noord, G. J.: 2009, Large Scale Syntactic Annotation of written Dutch (LASSY).
URL: *http://www.let.rug.nl/vannoord/Lassy/*

van Noord, G. J., Bouma, G., van Eynde, F., de Kok, D., van der Linde, J., Schuurman, I., Sang, E. T. K. and Vandeghinste, V.: 2013, Large scale syntactic annotation of written Dutch: LASSY, *Essential Speech and Language Technology for Dutch*, Theory and Applications of Natural Language Processing, Springer, pp. 231–254.

van Oosten, P., Tanghe, D. and Hoste, V.: 2010, Towards an improved methodology for automated readability prediction, *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC-2010)*, pp. 775–782.

Vapnik, V. and Cortes, C.: 1995, Support vector networks, *Machine Learning* **20**, 273–297.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D. and Hirschman, L.: 1995, A model-theoretic coreference scoring scheme, *Proceedings of the 6th Message Understanding Conference (MUC-6)*, pp. 45–52.

Vintar, v.: 2010, Bilingual term recognition revisited: The bag-of-equivalents term alignment approach and its evaluation, *Terminology* **16**, 141–158.

Vossen, P.: 1998, *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*, Kluwer, Dordrecht.

Vossen, P., Maks, I., Segers, R., van der Vliet, H., Moens, M., Hofmann, K., Sang, E. T. K. and de Rijke, M.: 2013, Cornetto: a lexical semantic database for Dutch, *in* P. Spyns and J. Odijk (eds), *Essential Speech and Language Technology for Dutch, Theory and Applications of Natural Language Processing*, Springer, pp. 165–184.

Webber, B.: 1978, *A Formal Approach to Discourse Anaphora*, PhD thesis, Harvard University.

Webber, B. and Joshi, A.: 2012, Discourse structure and computation: Past, present and future, *Proceedings of the Special Workshop on Rediscovering 50 Years of Discoveries (ACL-2012)*, pp. 42–54.

Wiebe, J., Wilson, T. and Cardie, C.: 2005, Annotating expressions of opinions and emotions in language, *Computer Intelligence* **39**(2), 165–210.

Wiegand, M. and Klakow, D.: 2010, Convolution kernels for opinion holder extraction, *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-2010)*, pp. 795–803.

Wilson, T., Wiebe, J. and Hoffman, P.: 2009, Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis, *Computational Linguistics* **35**(3), 399–433.

Wilson, T., Wiebe, J. and Hoffmann, P.: 2005, Recognizing contextual polarity in phrase-level sentiment analysis, *Proceedings of the 2005 Conference Empirical Methods in Natural Language Processing (EMNLP-2005)*, pp. 347–354.

Wolf, A.: 2005, Basic skills in the workplace: Opening doors to learning, *Technical report*, Chartered Institute of Personnel and Development.

Wright, S. E.: 1997, Term selection: the initial phase of terminology management, *in* S. E. Wright and G. Budin (eds), *Handbook of terminology management*, John Benjamins, pp. 13–23.

Yang, X., Su, J., Lang, J., Tan, L. C., Liu, T. and Li, S.: 2008, An entity-mention model for coreference resolution with inductive logic programming, *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL-2008)*, pp. 843–851.

Yang, X., Zhou, G., Su, S. and Tan, C.: 2003, Coreference resolution using competition learning approach, *Proceedings of the 41th Annual Meeting of the Association for Computational Linguistics (ACL-2003)*, pp. 176–183.

Zabin, J. and Jefferies, A.: 2008, Social media monitoring and analysis: Generating consumer insights from online conversation, *Technical report*, Aberdeen Group Benchmark Report, Aberdeen Group.

Zhang, Z., Iria, J., Brewster, C. and Ciravegna, F.: 2008, A comparative evaluation of term recognition algorithms, *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC-2008)*, pp. 2108–2113.

Zhu, J., Wang, H., Tsou, B. K. and Zhu, M.: 2009, Multi-aspect opinion polling from textual reviews, *Proceedings of the 18th Association for Computing Machinery Conference on Information and Knowledge Management (CIKM-2009)*, pp. 1799–1802.