Discourse lexicon induction for multiple languages and its use for gender profiling

Ben Verhoeven and Walter Daelemans
CLiPS Research Center, University of Antwerp, Belgium

Abstract

We propose a novel way to create categorized discourse lexicons for multiple languages. We combine information from the Penn Discourse Treebank with statistical machine translation techniques on the Europarl corpus. Using gender profiling as an application, we evaluate our approach by comparing it with an approach using features from a knowledge-based lexicon and with an Rhetorical structure theory (RST) discourse parser. Our experiments are performed on corpora for three languages (English, Dutch, and German) in two genres (news and blogs). We include a feature analysis in which we look for (in)consistencies of discourse features related to male and female authors between the different experimental settings.

Correspondence:
Walter Daelemans,
University of Antwerp,
Prinsstraat 13, 2000
Antwerp, Belgium.
E-mail: walter.daelemans@uantwerpen.be

1 Introduction

Computational discourse analysis is still quite limited by the number of languages for which discourse parsers or large enough resources exist. To our knowledge, there exists research on discourse parsing for only a handful of languages: English and Chinese—which were both part of the CoNLL-2016 shared task on shallow discourse parsing (Xue et al., 2016)—Brazilian Portuguese (Maziero et al., 2015; Pardo and Nunes, 2008), and those languages plus an additional four (Spanish, Dutch, Basque, and German) were recently studied by Braud et al. (2017). Not all of this research has led to practically useable discourse parsers though.

Recent initiatives such as the TextLink network have created an impetus for the creation of new resources (and for the unification of previously existing resources) related to discourse structure of text. Discourse-annotated corpora, as well as lexicons of discourse connectives, are now becoming available for an increasing number of languages. Discourse connectives are words or phrases that signal discourse relations (e.g. cause or contrast) between a sentence and what comes before or after. These resources are listed on the TextLink website,[1] where we count resources for more than fifteen different languages. Yet, several of these resources are very small and there is a lot of work to be done.

This article aims to contribute to the language diversity in discourse analysis by proposing a novel way to create discourse lexicons for multiple languages. Such lexicons contain discourse connectives that are annotated with the discourse relations they convey according to the Penn Discourse Treebank (PDTB) tags (The PDTB Research Group, 2008).

Our main interest is how the explicit discourse information in a text can be used as features—using such lexicons—for author profiling experiments, e.g. gender prediction where the task is to predict the gender of the author of a text, based on only the text. We believe that there are differences between individuals in how they convey the relations between sentences in a text, and also in which relations

Digital Scholarship in the Humanities

doi:10.1093/llc/fqy025

they use. For example, some people may make more comparisons than other people. It is our hypothesis that we can generalize and find discourse aspects of text that surpass the individuals and are found at the level of gender.

We perform this research on Dutch, English, and German corpora. English is interesting because we will be able to compare our approach with a discourse parser. We have included German in this study to be able to evaluate our approach extrinsically by comparing its performance with that of a knowledge-based discourse lexicon where connectives have associated relations, namely, DiMLex (Stede, 2002). We hypothesize that our approach will achieve similar performance to the knowledge-based lexicon, but that using the discourse parser will still outperform our approach.

The remainder of this article is structured as follows. We discuss some relevant related research on both discourse analysis and gender profiling in Section 2. We introduce Discourse for Multiple Languages (DiMuL)—our new approach to discourse features—as well as the other features we use in Section 3. In Section 4, we describe the data sets that we evaluate our features on. Our experiments are described in Section 5, in which you will also find the results. Section 6 contains a feature analysis. A discussion of all our findings follows in Section 7, after which a conclusion closes this article in Section 8.

2 Related Research

In this section, we provide a brief overview of related research on both discourse analysis and gender profiling. In a third subsection, we discuss the recent developments on the use of discourse features in gender profiling experiments.

2.1 Discourse analysis

Discourse is the level of information in text above that of the sentence. It is concerned with how text is structured and how text is coherent. 'A coherent text is designed around a common topic [...] individual units of information enter meaningful relationships to one another' (Stede, 2012, p. 1). One way to cohesively connect information over sentence boundaries is the use of connectives, such as 'in contrast' or 'moreover'. However, discourse relations between sentences can also exist implicitly, without the use of connectives. There is some research on the detection of these implicit discourse relations (Pitler et al., 2009; Lin et al., 2009), but it falls out of the scope of this article.

Discourse connectives (also known as discourse markers) are then important explicit markers of the discourse relations that are described in different theories of discourse structure. The best known models/resources of textual discourse are rhetorical structure theory (RST) (Mann and Thompson, 1987) and the PDTB (Prasad et al., 2008). In both cases, discourse connectives (sometimes implicit) get assigned discourse relations that denote the connection between two arguments, i.e. a sentence and what comes before or after. For example, Argument 2 can be the result of Argument 1, which may be indicated with the connective 'as a result'. We will discuss both the PDTB (see Section 3.1.1) and RST (see Section 3.2) further in this article. For a broader introduction into discourse analysis for language technology, see Webber et al. (2012) or Stede (2012).

2.2 Gender profiling

The goal of gender profiling is to predict the gender of the author of a text based only on the text itself. It is part of the broader task of author profiling where also other characteristics of the author are of interest, e.g. age or personality. Gender profiling is a well-established task that has received considerable attention over the years, not in the least because of online anonymity also creating harmful situations, such as sexually transgressive behavior. Since 2013, a yearly shared task on gender prediction has been organized as part of the PAN workshop series (Rangel et al., 2017). An overview of recent research in author profiling can be found in Neal et al. (2017).

A typical gender profiling system uses a supervised machine learning approach to classify texts into two classes, either male or female. Commonly used features include word and character *n*-grams.

In all previous research on gender profiling, gender is defined as a binary categorization between male and female. There is however an ongoing debate on non-binary gender[2] and the ethical implications around gender as a variable in natural language processing (NLP) research (Larson, 2017).

2.3 Discourse features for author profiling

In recent years, a few papers appeared that also investigated the importance of discourse features for author profiling. They will be briefly discussed here.

The first use of discourse features relevant to our research was not on author profiling but on authorship attribution, which is the task of attributing a text to a specific author rather than predicting the profile of an author. Feng and Hirst (2014) built a classifier on the work of nineteenth-century writers that predicts the author of a text. They use a method called discourse entity grid (Barzilay and Lapata, 2008) as a measure of discourse coherence. This method makes use of the different grammatical roles that the same entity can assume in a piece of text. The surface form of the entity can vary over the different occurrences. They hypothesize that authors individually differ in their use of patterns and types of role transitions. Recently, Ferracane et al. (2017) used the same featurization method in a new setup, employing a convolutional neural network for classification. They also proposed a method for the creation of discourse embeddings and found models using those to perform better than probability-based models.

To our knowledge, Soler-Company and Wanner (2017) were the first to study the relevance of discourse features for author profiling. They used the discourse parser by Surdeanu et al. (2015) to obtain the RST discourse structure of text. As features they use the frequency of each discourse relation and the shape of the discourse trees: depth, width, and ramification factor.[3] They found that discourse features are indeed informative of the age and gender of an author, though it is unclear whether it can also help to improve over state-of-the-art performance (Soler-Company, 2017).

3 Feature generation

In this section, we first propose our new approach DiMuL to generating discourse features that is applicable to many languages. We further introduce the discourse parser that we use for English. Finally, we explain how we created function word lexicons for the three languages in our study. Those lexicons are used to compare the performance of function word counts as features with the performance of discourse features in the context of gender prediction. All these features will be used in our experiments in addition to the standard features such as token and character *n*-grams.

3.1 Discourse for multiple languages

Our approach relies on both the PDTB and methods from statistical machine translation which we will discuss in the following sections. We first describe an example of our approach: the transfer of the discourse relations of an English connective to a different language.

3.1.1 *Penn Discourse Treebank*

The PDTB is an English text corpus in which occurrences of discourse connectives are annotated with the discourse relation they convey according to the PDTB 2.0 tag set (Prasad et al., 2008). Depending on the context, the same connective can entail different discourse relations. There are 100 unique connectives present in the corpus, and we obtain all the annotations for all their occurrences. This allows us to compute frequencies of discourse relations for each connective which can later be used as weights. The PDTB tags are hierarchically structured, which enables us to create discourse lexicons on three levels of specificity (from four very broad relations to forty very specific relations). Figure 1 shows the tag set and illustrates its hierarchy.

Example—Part 1

The English discourse connective *moreover* has expansion as associated PDTB relation with 1.0 as strength, because it apparently only occurs with this meaning in the PDTB corpus.

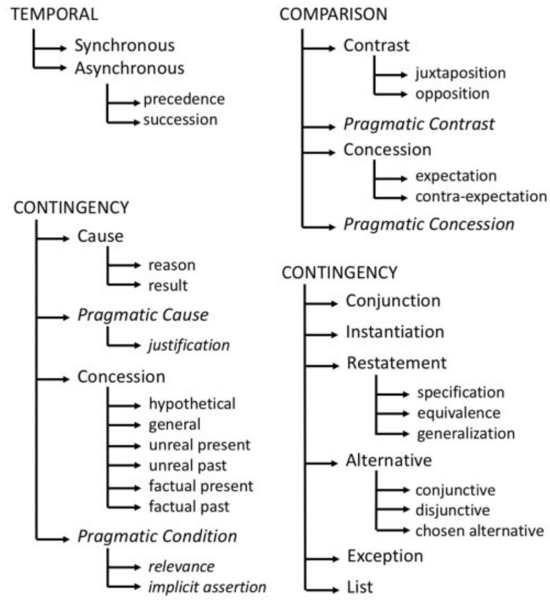

Fig. 1 Relation hierarchy of the PDTB 2.0 tagset (The PDTB Research Group, 2008, p.27)

3.1.2 *Extrapolate to other languages*

The multilingual aspect of our approach relies on methods from statistical machine translation (Brown et al., 1990). By measuring how frequently words and phrases appear as translations of each other in parallel corpora, phrase tables can be made that list elements with the strongest translational equivalence, i.e. words that are each other's best translations (Melamed, 2000).

Lopes et al. (2015) produced such phrase tables of discourse connectives for all the languages in the Europarl parallel corpus (Koehn, 2005). They are available online,[4] and we will use them in this study. The novelty of our work is in extrapolating the English-weighted discourse lexicon from the PDTB to other languages (such as Dutch and German), using these phrase tables.

We hypothesize that abstracting over the individual connectives to discourse relations can provide stronger evidence of the use of these discourse relations. For example, in Dutch, *als* and *indien* (English: *if*) are both markers that indicate a condition. Instead of as two separate words we will count them as two instances of the same relation. Furthermore, studying the relations apart from the connectives allows us to make comparisons of discourse between languages.

Because we utilize phrase tables from the Europarl corpus (Koehn, 2005), we can easily create our weighted discourse lexicons for all official languages of the European Union.

The English list of discourse connectives used by Lopes et al. contains 456 items, while the PDTB contains 100 different discourse connectives. There is an overlap of 86 items between the two lists that we can use as our seed list. The remaining items are thus discarded because we have no associated discourse relations for them. After extrapolation from English using the phrase tables, we find 335 unique connectives for Dutch and 341 unique connectives for German. The mean number of discourse relations per connective is around 1.9. The mean strength of the strongest relation of each connective is 0.90, while the median strength is 0.98. This indicates that although a connective has two associated relations on average, most connectives have a strong primary meaning.

To evaluate our approach, we need an existing knowledge-based resource similar to our weighted lexicon. If we can show our discourse features to perform at least as well as those generated with the knowledge-based lexicon, we can assume that our approach is solid. We found the German discourse lexicon DiMLex to fit our needs perfectly. DiMLex is a lexicon of German discourse markers with manually annotated relations of discourse per connective (Stede, 2002; Scheffler and Stede, 2016). We will discuss the outcome of this evaluation in Section 7. Since there do not exist such lexicons for English and Dutch, we will have to assume that this approach—if successful—also works for other languages.

Example—Part 2

We find the following five Dutch collocations using Lopes et al.'s phrase tables with *moreover* as the source word: *trouwens, verder, voorts, bovendien de, bovendien in*. Each of these Dutch target connectives now receives the discourse relations and their weights of the source word, namely, expansion with weight 1.0. Unless they have multiple source words (when the connective is the target word of more than one source word), then the relations of the source words are merged with weighting by collocation strength. In our example, *trouwens* has a second source word: *besides*. This word has two associated relations in the PDTB: expansion (weight: 0.9474) and comparison (weight: 0.0526). The strength of the collocation of the source words with the target words (*moreover*: 0.1068; *besides*: 0.0037) now serves as a weight for their relations to be combined to the target relations. We take the product of the collocation strength and the source relations. After rescaling to sum is 1, the target relations become expansion (weight: 0.9982) and comparison (weight: 0.0018).

3.1.3 *DiMuL features*

Featurizing a text with the DiMuL lexicons works as follows. The occurrences of each connective in the lexicon are counted. In the case of multi-word phrases, only the longest matching connective is counted. We then assume that the associated relations of each connective are present in the proportion of their weights. Some of our discourse connectives also have non-connective uses. Just like a bag-of-words model, we do not distinguish between different meanings of words or phrases.

3.2 RST parser

We compare our discourse features on the English data with features from the RST discourse parser by Surdeanu et al. (2015) available on Github[5] which was also used by Soler-Company and Wanner (2017). We employ a similar method of featurizing the discourse parser output as in this previous research, namely, use counts of the identified discourse relations normalized by the document length.

These features are actually very similar in design to our discourse features, but where the RST parser processes an entire text, our features are generated based only on token cues in the text. We can distinguish two levels of specificity by either taking the direction of the RST relation into account or not, for example contrast (RightToLeft) vs. contrast. The higher specificity stands for the more specific relations. Where other researchers sometimes also add features representing the depth and width of the RST parse tree, we decided not to use those to keep the approach comparable to ours and focus on information about the types of discourse relation.

3.3 Function words

As a point of comparison for our discourse approach, we will also do experiments with function word counts as features. Function words are considered standard features for gender prediction experiments. We gathered the following lists of function words for the three languages under consideration. These lists were compiled in different ways—as described below—because of which we cannot compare the performance of function words over the different languages, but it does allow us to compare with the performance of other features for the same language.

English: We used the list[6] of 277 function words from O'Shea et al. (2010). This list was compiled 'by combining stop word lists, removing the content words and then adding low frequency function words from dictionaries' (O'Shea et al., 2010).

Dutch: We manually selected the most frequent 450 function words from the SUBTLEX word frequency list[7] compiled by the Centre for Reading Research from Ghent University.

German: We manually compiled a list of function words that were found on two websites.[8] This list contains 145 function words.

4 Data sets

We use a total of five data sets for three languages. The size and class distribution of each corpus can be found in Table 1. In the following sections, we describe the origins of each corpus and how we processed it. In light of the focus of the article on discourse elements, we chose two genres that typically do not have very short texts (such as typical social media text might have), namely, news articles and blog posts.

4.1 News corpora

Two existing news corpora were newly annotated for gender. We have a Dutch corpus from the

Table 1 Table of data sets and their total and per-gender sizes in number of texts/instances

Corpus	Language	Genre	Female	Male	Total
NYT	English	News	155,219	153,916	309,135
HLN	Dutch	News	72,161	72,163	144,324
Blogger-NL	Dutch	Blogs	84,363	84,363	168,726
Blogger-DE	German	Blogs	238,105	238,105	476,210
Blog Authorship	English	Blogs	308,358	308,358	616,716

Flemish-Belgian newspaper *Het Laatste Nieuws* (*HLN*) at our disposal, as well as an English corpus from the *New York Times* (*NYT*). These corpora have in common that the name of the journalist is present in the metadata which allows for deriving the gender of the author. This gender annotation was done automatically with manual verification. Only the authors of single-author articles were taken into account, whose names were then automatically classified as male or female using a gazetteer-based approach[9] and manually verified. In case of doubt—for example with ambiguous names such as Alex, Sam, Kim and others—the journalist was discarded.

Both corpora were further processed to create instances in such a way that there was a balance for gender. Also, a single author is not allowed to have more than 1,000 articles in the final data set.

In the next two subsections, we provide some more details about the two news corpora we used. Unfortunately, we did not have a German news corpus to our disposal.

4.1.1 *Het Laatste Nieuws*

This corpus exists of news articles published in the period 2010–16 in the Belgian Dutch-language newspaper *HLN*. The result of our gender annotation is a list of 699 authors of which 547 are men and 152 are women. One million articles are associated with these authors. At the end of the preprocessing, we end up with 144,324 instances, balanced for gender.

4.1.2 *New York Times*

This corpus contains news articles that were published in the *NYT* in the period 1987–2006. The result of the gender annotation is a list of 3,579 authors of which 2,174 are men and 1,405 are women. Almost one million articles are associated with these authors. At the end of the preprocessing, we end up with 309,135 instances, balanced for gender with 153,916 articles by men, and 155,219 articles by women.

4.2 Blog corpora

The second genre is blogs. For English, we use the Blog Authorship corpus that was developed by Schler et al. (2006). We also use a Dutch and German corpus that originates from the Blogger platform.

4.2.1 *Blogger*

In 2014, some large corpora of blog posts in different languages were created in our research group by mining[10] the Blogger[11] platform for blog posts. Many bloggers list their gender on their profile which provides us with a gold standard label. We will employ the German and Dutch corpora in this research while balancing the data sets for gender. This results in a Dutch data set of 168,726 instances and a German data set of 476,210 instances.

4.2.2 *Blog Authorship Corpus*

The Blog Authorship corpus is an English corpus of blog posts. It was published in 2006 by Schler et al. and is still available online.[12] The data were collected in 2004, also from the Blogger platform but because there is a huge time gap and this corpus has been published, we describe it separately. The age and gender of each blogger are known, as well as the astrological sign and industry of some bloggers.

The corpus was preprocessed by us—discarding some files with problematic encoding—and balanced for gender, leaving us with 616,716 instances.

5 Experiments

We run standard supervised machine learning experiments using the Python libraries omesa[13] for vectorization and sklearn (Pedregosa et al., 2011) for classification. We exploratively test a number of different classification algorithms in a ten-fold cross-validation setup. We work with algorithms that have provided good results in related research.

Table 2 All abbreviations and their explanations

Abbreviation	Explanation
SGD	Stochastic Gradient Descent classification
LR	Logistic Regression classification
RF	Random Forest classification
ML	Machine Learning
baseline	Majority baseline, independent of algorithm
tok1	Token unigrams
tok2	Token bigrams
char3	Character trigrams
char4	Character tetragrams
dimulcat1	DiMuL discourse relations with specificity 1
dimulcat2	DiMuL discourse relations with specificity 2
dimulcat3	DiMuL discourse relations with specificity 3
dimulconn	DiMuL discourse connectives
funcwords	Function words
dimlexcat1	DiMLex discourse relations with specificity 1
dimlexcat2	DiMLex discourse relations with specificity 2
dimlexcat3	DiMLex discourse relations with specificity 3
dimlexconn	DiMLex discourse connectives[15]
rst1	Features of the RST parser with specificity 1
rst2	Features of the RST parser with specificity 2

Whenever possible, the parameter random_state (which controls the internal randomization of the algorithms) was fixed to be able to reproduce results. We used the following machine learning algorithms:

- SGDClassifier[14] with $n\_\ iter = 50$
- Logistic Regression (LR)
- Random Forest (RF) Classifier

In the following two sections we will describe the results of our gender prediction experiments. Table 2 provides a summary of all the abbreviations used in the following result tables. The baseline for all corpora is 0.50, since we are dealing with a two-class problem and they are all balanced. We have run experiments on each feature type separately, as well as on combinations of *n*-gram features and discourse features. The discourse features never improved the *n*-gram results.

5.1 Results for news

The results for both news corpora show moderate results for state-of-the-art features, such as token and character *n*-grams. For Dutch, the results are around 0.63–0.64 in *F*-score (see Table 3). For English, the results (see Table 4) are slightly higher with *F*-scores between 0.66 and 0.68.

Table 3 Results in *F*-score for gender classification on the **Dutch HLN** data set using different ML algorithms

Feature type	SGD	LR	RF	# Features
tok1	0.63	0.63	0.55	1,283,954
tok2	0.62	0.63	0.56	8,021,660
char3	0.62	0.64	0.58	115,921
char4	0.64	0.64	0.58	657,989
dimulcat1	0.49	0.46	0.51	4
dimulcat2	0.48	0.45	0.52	20
dimulcat3	0.49	0.47	**0.51**	40
dimulconn	0.52	0.49	**0.52**	335
funcwords	0.50	0.51	0.52	450
baseline		0.50		

Table 4 Results in *F*-score for gender classification on the **English NYT** data set using different ML algorithms

Feature type	SGD	LR	RF	# Features
tok1	0.66	0.67	0.57	5,438,688
tok2	0.68	0.68	0.57	47,393,500
char3	0.63	0.66	0.56	129,107
char4	0.66	0.68	0.56	939,064
dimulcat1	0.48	0.51	0.50	4
dimulcat2	0.48	0.53	0.51	20
dimulcat3	0.50	**0.54**	0.51	40
dimulconn	0.51	**0.54**	0.51	100
funcwords	0.53	0.58	0.54	277
rst1	0.34	0.35	0.50	18
rst2	0.34	0.56	0.54	42
baseline		0.50		

When looking at our discourse features, it is harder to make a clear analysis. The results for discourse on the Dutch HLN corpus do not seem to outperform the baseline. For English, there seems to be a slightly larger learning effect but still quite close to the baseline. The RST features with higher specificity seem to outperform our features. The lower specificity RST features do not perform well.

Combinations of *n*-gram features with discourse features were empirically tested for both the news and blogs corpora, but the results were always near-identical to the *n*-gram result.

5.2 Results for blogs

We achieve very high results with *n*-gram features for both the German (0.88–0.92, see Table 7) and Dutch (0.89–0.92, see Table 5) Blogger data sets.

Table 5 Results in *F*-score for gender classification on the **Dutch Blogger** data set using different ML algorithms

Feature type	SGD	LR	RF	# Features
tok1	0.91	0.91	0.79	2,940,191
tok2	0.89	0.89	0.78	14,404,524
char3	0.90	0.90	0.78	216,435
char4	0.91	0.92	0.80	1,182,147
dimulcat1	0.54	0.59	0.57	4
dimulcat2	0.55	0.63	0.61	20
dimulcat3	0.57	**0.63**	0.61	40
dimulconn	0.60	**0.68**	0.65	335
funcwords	0.80	0.81	0.77	450
baseline		0.50		

The English Blog Authorship Corpus yields somewhat more moderate results, with *F*-scores between 0.67 and 0.69 on *n*-gram features (see Table 6).

With regard to the discourse features, we again see very promising results for Dutch (0.59–0.68) and German (0.61–0.64). The table with the results for the German Blogger dataset also contains results for the DiMLex features, which score lower (0.58–0.61) than our own discourse features. For English, the discourse results are more comparable to the news corpus with a range of 0.53–0.54. Also similar to the news corpus is that the RST features outperform our discourse features (only) when using the more specific relations.

Table 6 Results in *F*-score for gender classification on the **English Blog Authorship Corpus** using different ML algorithms

Feature type	SGD	LR	RF	# Features
tok1	0.64	0.67	0.60	3,168,651
tok2	0.66	0.67	0.60	23,804,484
char3	0.65	0.68	0.59	501,662
char4	0.68	0.69	0.60	2,088,514
dimulcat1	0.46	0.50	0.53	4
dimulcat2	0.51	0.53	0.54	20
dimulcat3	0.50	**0.54**	0.54	40
dimulconn	0.50	**0.55**	0.54	100
funcwords	0.57	0.59	0.58	277
rst1	0.49	0.50	0.49	18
rst2	0.53	0.56	0.55	42
baseline		0.50		

Table 7 Results in *F*-score for gender classification on the **German Blogger** data set using different ML algorithms

Feature type	SGD	LR	RF	# Features
tok1	0.90	0.88	0.81	6,630,650
tok2	0.88	0.89	0.79	29,765,260
char3	0.89	0.89	0.80	1,023,432
char4	0.92	0.92	0.82	3,510,414
dimulcat1	0.44	0.49	0.61	4
dimulcat2	0.48	0.52	0.62	20
dimulcat3	0.49	0.53	**0.62**	40
dimulconn	0.54	0.57	**0.64**	341
dimlexcat1	0.43	0.47	0.58	4
dimlexcat2	0.42	0.48	0.59	11
dimlexcat3	0.36	0.48	0.60	18
dimlexconn	0.46	0.51	0.61	59
funcwords	0.69	0.71	0.75	143
baseline		0.50		

6 Feature Analysis

In this section we perform a feature analysis aimed to investigate which discourse relations and connectives are good predictors of either female or male authorship. We can use the coefficients of the LR learning algorithm to identify good features. Features with positive coefficients are associated with the positive class and vice versa. The higher the absolute value of the coefficient, the stronger the association with the class. It is difficult to draw any conclusions for Dutch, since the classifier trained on the *HLN* Dutch news corpus with discourse features has a very weak performance.

In a first analysis, we have a look at the ten connectives that are associated the strongest with each gender for each of the corpora (except the *HLN* Dutch news, see above). We can say the most about the English connectives (see Table 8) because we have two corpora available for this language. We see that *in other words* and *in short* are considered male connectives in both corpora and that *when and if* is considered female in both corpora (see **bold** marking). The connectives for German can be found in Table 9, while the Dutch connectives are listed in Table 10. For German, two connectives for which the translation is also in the English top ten behave conversely. The connectives *nichtsdestoweniger* and *mithin* are associated with the male gender in German, but *nonetheless* and *consequently* are

Table 8 Ten connectives with strongest association with male and female gender for the English corpora: *NYT* (news) and Blog Authorship Corpus (blogs)

Male		Female	
News	**Blog**	**News**	**Blog**
if then	conversely	neither nor	either or
insofar as	**in short**	**when and if**	separately
thereby	as an alternative	additionally	by then
accordingly	overall	if and when	**when and if**
thereafter	ultimately	alternatively	now that
moreover	**in other words**	nonetheless	until
in other words	whereas	before and after	consequently
in short	as though	for instance	when
indeed	for instance	although	because
by then	in the end	as well	besides

Note. The words are ranked with the strongest association on top. Connectives that occur in the top ten of both corpora for the same gender are indicated in **bold**.

Table 9 Ten connectives with strongest association with male and female gender for the German Blogger corpus and their English translations

Male		Female	
German	**Translation**	**German**	**Translation**
faktisch	factually	sich vorhin	themselves recently
als resultat	as a result	der für	who for
nichtsdestoweniger	nonetheless	die speziell	who especially
vormals	formerly	so bald wie	as soon as
gleichfalls	also	außerdem	in addition
infolgdessen	as a result	zwischenzeit	meantime
die insbesondere	who especially	gleichwohl der	as well as the
gesamtheit	entirety	so sehr	so much
mithin	consequently	die daher	who as a result
der insbesondere	who especially	sehr wie	so much as

Note. The words are ranked with the strongest association on top. Connectives whose translations occur in the top ten for English for the opposite gender are indicated in *italics*.

both associated with female gender in English. For Dutch, we were able to spot two connectives behaving conversely to the other languages. The connectives *andere woorden* and *als resultaat* are associated with the female gender in Dutch, while their translations *in other words* (English) and *als resultat* (*as a result* in German) are associated with the male gender.

Our second analysis is based on the dimulcat3 features for all corpora which are our own DiMuL features at specificity Level 3 (most specific). The most important relation in our analysis seems to

Table 10 Ten connectives with strongest association with male and female gender for the Dutch Blogger corpus and their English translations

Male		Female	
Dutch	**Translation**	**Dutch**	**Translation**
evenwel	however	is zodat	is so that
evenzeer	likewise	*andere woorden*	other words
als geheel	as a whole	*als resultaat*	as a result
aldus	thus	komen later	come later
daarom te	therefore	zo van	like
de sinds	since	nochtans	nevertheless
onderwijl	during	eindelijk van	at last
voorts	furthermore	zolang in	as long in
dat uiteindelijk	that eventually	tot slot	finally
desondanks	despite that	zodra de	as soon as the

Note. The words are ranked with the strongest association on top. Connectives whose translations occur in the top ten of a different language for the opposite gender are indicated in *italics*.

Table 11 Features linked to genders per language over different genres (blogs and news)

	English	**Dutch**	**German**
Female		*Concession*	*Concession*
	Pragmatic contrast	Opposition	*Unreal past*
	Implicit assertion		*Unreal present*
Male	Unreal present	**Unreal present**	
	Temporal		**Temporal**
	Contingency		
	Factual present		
	Unreal past		
	Concession		

be justification which behaves interestingly with regards to genre. This relation is strongly male in all the news corpora, yet strongly female in all the blogs corpora. The discourse relation of justification entails that a claim is being expressed and that some justification for this claim is present (though no causal influence is implied) (The PDTB Research Group, 2008, p. 29). We currently have no adequate explanation for this finding.

We also investigate our features within each language over the two genres. We list the most important relations in Table 11 below and indicate where Dutch and German are similar to (**bold**) or different from (*italics*) English. We will focus our analysis on German and English. These languages agree on one relation: temporal seems to be strongly male. The other three important relations for German (concession, unreal past, and unreal present) are female in German, yet male in English.

7 Discussion

Our findings show that discourse aspects of a text contribute to the prediction of the gender of a text's author. There are however several interesting remarks to be made about the circumstances in which this is possible.

The results are much better for the Blogger data sets (Dutch and German) than for the news data sets (*HLN* for Dutch and *NYT* for English). This may indicate that there is a genre factor to gender prediction from discourse. News is a more formal genre in which authors follow editing guidelines, and their articles are often edited. This may reduce the impact of the individual writing style (containing gender elements) on the text. Strangely, the results on the English blogs are in between the results for news and for the other blogs. There is no easy explanation for this, though we want to note that this particular corpus was constructed in the early 2000s when blogs as the genre we know now was still in its infancy.

The algorithm that performs best overall is LR. Though stochastic gradient descent (SGD) is very fast and performs well for experiments with many features, it performs badly when there are few features. The experiments with RF show especially good results on the experiments for German Blogger with few features. Early on in the study, we also tested MultinomialNB and DecisionTreeClassifier and tried out scaling and different ways of performing feature selection. However there was no noticeable speedup nor improved results which left us with no reason to use them further.

Looking at which feature types perform best, there seems to be a trend that the more specific the features a system can use, the better the system. This is very visible in the following hierarchy of features, ordered from lower to better results, which is consistent over all the data sets: dimulcat1 < dimulcat2 < dimulcat3 < dimulconn < funcwords. A similar remark was made by Ferracane et al. (2017, p. 8) when discussing the performance of discourse embeddings which also allow for more specific features. Unfortunately, one negative finding is that discourse relations do not perform better than the connectives.

There are two comparisons to be made to evaluate our approach. We used the approximate randomization test (ART)[16] to find out whether the systems were significantly different from each other. ART is a nonparametric test suitable for *F*-scores (Noreen, 1989).

We first tested dimulcat3 against dimlexcat3 on the German Blogger data set ($P < 0.001$). Interestingly, our generated lexicon DiMuL thus performs significantly better than the knowledge-based lexicon DimLex (which can be considered a gold standard) on the German data set for gender detection. This is better than we hypothesized. In light of our previous finding (see paragraph above), this is probably due to the larger number of (more specific) features that DiMuL can create, but that is also an advantage of our approach.

Second, we tested rst2 versus dimulcat3 on the English NYT corpus: ($P < 0.001$) and found the RST discourse parser to significantly outperform our own approach. Since our approach is an approximation by weighting of the discourse relations in the text based on the connectives, and the RST parser constructs a discourse parse tree of the entire text, this result is not surprising, as the parser is expected to be more precise than our system. However, such parsers only exist for a very limited number of languages, thus maintaining the relevance of our approach.

8 Conclusion and Future Work

In summary, we have presented a novel approach (DiMuL) for inducing lexicons of discourse connectives with associated discourse relations by combining an existing English corpus of discourse (PDTB) with techniques from machine translation. We have created resources for Dutch and German, but because we use the Europarl corpus, our method will work for all European languages.

Furthermore, we have used these lexicons for gender prediction experiments. The results vary between the data sets and genres. The results for the news corpora were inconclusive, but the blog corpora showed convincing predictive effects of discourse aspects of text. Using the connectives as features gave better results than using the discourse relations as features, most likely due to the reduced number of features. Our approach gave better results than using a knowledge-based lexicon for German; yet the available English discourse parser still outperformed our approach.

As this is still one of the first works on utilizing discourse information for author profiling, there is much work to be done. We are strong proponents of research on less prominent languages, so we hope that other researchers will employ our method to explore what discourse lexicons can offer for their languages. Also, we have chosen for this article not to work with implicit discourse relations, but we believe this to be an important line in future research. For example, because there might be differences between groups of people in how many (and which) relations they express explicitly versus implicitly.

Notes

1 http://textlink.ii.metu.edu.tr/
2 Using non-binary gender is currently still unfeasible for NLP research due to lack of data, but we strive in this article to be as transparent as possible about the origin of our gender labels in each of the corpora.
3 he ramification factor is the mean number of children nodes per level of the discourse tree.
4 https://goo.gl/3jicxV
5 We use the dependency parser in FastNLPProcessor at https://github.com/clulab/processors
6 https://semanticsimilarity.wordpress.com/function-word-lists/
7 http://crr.ugent.be/programs-data/subtitle-frequencies/subtlex-nl/downloading
8 http://www.lingudora.com/en/learn-german-online/vocabulary/list/5 and http://www.vistawide.com/german/top_100_german_words.htm
9 We used the Textgain gender API for this purpose: https://www.textgain.com/api#gender
10 https://bitbucket.org/enrique_manjavacas/blogproj/
11 http://www.blogger.com
12 http://u.cs.biu.ac.il/~koppel/BlogCorpus.htm
13 https://github.com/cmry/omesa
14 For SGD to work properly, the n_iter parameter should have a minimum total of 1 million over all instances. See: http://scikit-learn.org/stable/modules/sgd.html#tips-on-practical-use
15 The number of features listed in the result tables depends on the number of connectives actually occurring in the corpus.
16 We used the implementation by Vincent Van Asch: https://www.clips.uantwerpen.be/scripts/art

Acknowledgements

The authors would like to thank the following of our colleagues: Enrique Manjavacas for collecting the Blogger corpora, Simon Šuster for help with the RST parser, and Giovanni Cassani for useful insights on statistical analyses that did not make it into the final version of the article. B.V. is supported by a PhD scholarship from the FWO Research Foundation—Flanders.

References

Barzilay, R. and Lapata, M. (2008). Modeling local coherence: an entity-based approach. *Computational Linguistics*, **34**(1): 1–34.

Braud, C., Coavoux, M., and Søgaard, A. (2017). Cross-lingual RST Discourse Parsing. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Valencia, Spain: ACL, pp. 292–304.

Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. (1990). A statistical approach to machine translation. *Computational Linguistics*, **16**(2): 79–85.

Feng, V. W. and Hirst, G. (2014). Patterns of local discourse coherence as a feature for authorship attribution. *Literary and Linguistic Computing*, **29**(2): 191–8.

Ferracane, E., Wang, S., and Mooney, R. J. (2017). Leveraging Discourse Information Effectively for Authorship Attribution. In Proceedings of the 8th International Joint Conference on Natural Language Processing. Taipei, Taiwan: ACL, pp. 584–93.

Koehn, P. (2005). Europarl: a parallel corpus for statistical machine translation. In *MT summit*, vol. **5**). Thailand: Phuket, pp. 79–86.

Larson, B. N. (2017). Gender as a variable in natural-language processing: ethical considerations. In

Workshop on Ethics in Natural Language Processing. Valencia, Spain: ACL, pp. 1–11.

Lin, Z., Kan, M.-y., and Ng, H. T. (2009). Recognizing Implicit Discourse Relations in the Penn Discourse Treebank. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, number August, Singapore, pp. 343–51.

Lopes, A., de Matos, D. M., Cabarrão, V., Ribeiro, R., Moniz, H., Trancoso, I., and Mata, A. I. (2015). Towards using machine translation techniques to induce multilingual lexica of discourse markers. *arXiv*, 1503.09144v1.

Mann, W. C. and Thompson, S. A. (1987). Rhetorical structure theory: a theory of text organization. Technical report, Technical Report ISI/RS 87-190, ISI.

Maziero, E. G., Hirst, G., and Pardo, T. A. S. (2015). Adaptation of Discourse Parsing Models for the Portuguese Language. In Proceedings - 2015 Brazilian Conference on Intelligent Systems, BRACIS 2015, pp. 140–5.

Melamed, I. D. (2000). Models of Translational Equivalence among Words. *Computational Linguistics*, **26**(2): 221–49.

Neal, T., Sundararajan, K., Fatima, A., Yan, Y., Xiang, Y., and Woodard, D. (2017). Surveying stylometry techniques and applications. *ACM Computing Surveys*, **50**(6).

Noreen, E. W. (1989). *Computer-intensive Methods for Testing Hypotheses: An Introduction*. Wiley, Hoboken, New Jersey, USA.

O'Shea, J. (2010). A Framework for Applying Short Text Semantic Similarity in Goal-Oriented Conversational Agents. PhD thesis, Manchester Metropolitan University, Manchester.

O'Shea, J., Bandar, Z., and Crockett, K. (2010). A Machine Learning Approach to Speech Act Classification Using Function Words. In *KES-AMSTA'10 Proceedings of the 4th KES international conference on Agent* and Multi-Agent Systems: Technologies and Applications. Heidelberg: Springer, pp. 82–91.

Pardo, T. A. S. and Nunes, M. d. G. V. (2008). On the development and evaluation of a brazilian portuguese discourse parser. *Revista de Informática Teórica e Aplicada*, **15**(2): 43–64.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, **12**: 2825–30.

Pitler, E., Louis, A., and Nenkova, A. (2009). Automatic Sense Prediction for Implicit Discourse Relations in Text. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, vol. 2. Suntec, Singapore: ACL; AFNLP, pp. 683–91.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008). The Penn Discourse TreeBank 2.0. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). Marrakech, Morocco: ELRA, pp. 1–4.

Rangel, F., Rosso, P., Potthast, M., and Stein, B. (2017). Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. In CEUR Workshop Proceedings. CEUR-ws.org.

Scheffler, T. and Stede, M. (2016). Adding Semantic Relations to a Large-Coverage Connective Lexicon of German. In Proceedings of LREC. Portorož, Slovenia: ELRA, pp. 1008–13.

Schler, J., Koppel, M., Argamon, S., and Pennebaker, J. W. (2006). Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, vol. **6**, AAAI, Palo Alto, California, USA, pp. 199–205.

Soler-Company, J. (2017). Feature Engineering for Author Profiling and Identification: On the Relevance of Syntax and Discourse. PhD thesis, Universitat Pompeu Fabra, Barcelona.

Soler-Company, J. and Wanner, L. (2017). On the Relevance of Syntactic and Discourse Features for Author Profiling and Identification. In 15th Conference of the European Chapter of the Association for Computational Linguistics. Valencia, Spain: ACL, pp. 681–7.

Stede, M. (2002). DiMLex: A lexical approach to discourse markers. In Lenci, A. and Di Tomaso, V. (eds), *Exploring the Lexicon—Theory and Computation*. Alessandria, Italy: Edizioni dell'Orso, p. 15.

Stede, M. (2012). Discourse Processing. In *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers.

Surdeanu, M., Hicks, T., and Valenzuela-esc, M. A. (2015). Two Practical Rhetorical Structure Theory Parsers. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies: Software Demonstrations (NAACL HLT). Denver, CO: ACL, pp. 1–5.

The PDTB Research Group (2008). *The Penn Discourse Treebank 2.0 Annotation Manual*. IRCS Technical Reports Series. The PDTB Research Group.

Webber, B., Egg, M., and Kordoni, V. (2012). Discourse structure and language technology. *Natural Language Engineering*, **18**(4): 437–90.

Xue, N., Ng, H. T., Pradhan, S., Webber, B., Rutherford, A., Wang, C., and Wang, H. (2016). The CoNLL-2016 Shared Task on Multilingual Shallow Discourse Parsing. In Proceedings of the Twentieth Conference on Computational Natural Language Learning - Shared Task. Berlin, Germany: ACL, pp. 1–19.