Identification of Differences between Dutch Language Varieties with the VarDial2018 Dutch-Flemish Subtitle Data

Hans van Halteren
Centre for Language Studies
Radboud University Nijmegen
P.O. Box 9103, 6500-HD Nijmegen
The Netherlands
hvh@let.ru.nl

Nelleke Oostdijk
Centre for Language Studies
Radboud University Nijmegen
P.O. Box 9103, 6500-HD Nijmegen
The Netherlands
N.Oostdijk@let.ru.nl

Abstract

With the goal of discovering differences between Belgian and Netherlandic Dutch, we participated as Team Taurus in the Dutch-Flemish Subtitles task of VarDial2018. We used a rather simple marker-based method, but with a wide range of features, including lexical, lexico-syntactic and syntactic ones, and achieved a second position in the ranking. Inspection of highly distinguishing features did point towards differences between the two language varieties, but because of the nature of the experimental data, we have to treat our observations as very tentative and in need of further investigation.

1 Introduction

The main area where the Dutch Language is spoken is in The Netherlands and the Northern part of Belgium (Flanders). Although there are quite strong dialects in regions in both countries, the standard version of Dutch is shared. In fact, there is a joint Dutch Language Union that promotes and supports standard Dutch. Still, many native speakers have the feeling that there are subtle differences between the Northern and the Southern variety of standard Dutch, and that these differences are not limited to just pronunciation. We would like to verify whether this intuition is correct, by investigating (qualitatively and quantitatively) the language use in corpus material that is meant to represent standard Dutch (and not, e.g., the language used on social media as that tends to contain high levels of dialect in various regions) and that is balanced in all factors apart from the language variety. However, such a corpus is hard to come by. The Spoken Dutch Corpus (Oostdijk et al., 2000) contains both varieties, but the content is necessarily biased by location. The same can be said about SoNaR (Stevin Dutch Reference Corpus; Oostdijk et al., 2013). We were therefore pleased to find that one of the tasks of VarDial 2018 (Zampieri et al., 2018), namely Discriminating between Dutch and Flemish in Subtitles (DFS), appeared to provide exactly what we were looking for: a corpus of subtitles of international movies and tv shows, produced by the Dutch and Belgian branches of Broadcast Text International (BTI Studios) and decided to participate, as Team Taurus.

As can be deduced from the introduction above, our main goal is not the highest possible recognition score; rather, we want to establish if the two varieties differ from each other and if so, in what respect. Of course, a score higher than chance is required to show that indeed there are differences between the varieties, but mostly we are interested in which features are apparently used in distinguishing between Northern and Southern Dutch. This means that we are limited in our choice of recognition methods. For example Support Vector Machines, although very strong in recognition quality, are not suited for our purpose as the transformation of the feature space makes evaluation per feature impossible. Instead, we chose a very simple marker-based method, which allows us to see directly how much each feature contributes. As for recognition features, our main interest lies in syntax, even for text classification already more than two decades (Baayen et al., 1996). Still, we chose as wide a range as we could extract in the time allotted to this project, ranging from character n-grams to syntactic rewrites.

Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 199–209
Santa Fe, New Mexico, USA, August 20, 2018.

In the sections below we will first describe some related work (Section 2) and then the experimental data and our preprocessing (Section 3). Next we describe our features in more detail (Section 4). Then we proceed with the recognition method and the recognition quality (Section 5). Our investigation on variety-distinguishing features in this paper will be restricted to token unigrams (Section 6) and syntactic features (Section 7). We conclude with a more general discussion of the results (Section 8).

2 Related Work

Existing work that is related to this paper can belong to several classes. Seeing the volume that this paper is published in, the most obvious class is also the one that needs least discussion: dialect recognition in general. Overviews of the field can be found in the VarDial reports (Zampieri et al., 2017; Zampieri et al., 2018). The field is dominated by text classification methods using knowledge-poor features, namely character n-grams and word n-grams. Differences are generally present in the choice of machine learning method and tuning approaches. Van der Lee and van den Bosch (2017) deserve a special mention, on various grounds. They compare Belgian and Netherlandic Dutch, using the SUBTIEL corpus from which the VarDial2018-DFS data has also been extracted. Furthermore, they explore a wide range of machine learning methods, as well as features based on POS-tagging (which they call 'syntactic features', whereas we reserve this term for features based on full syntactic analysis).

An entirely different field is that of linguistic studies into language variation, within which we now want to focus on differences between the Belgian and Netherlandic varieties of Standard Dutch (ignoring the many regional and local dialects that exist). Although many people seem to think that the two varieties are virtually identical in word use, and merely differ in pronunciation, native speakers do "feel" differences even in written texts. These intuitions also made their way into NLP, e.g. Despres et al. (2009) decided for their speech recognition of broadcast news not only to make specific acoustic models for the Dutch and Flemish datasets, but also separate lexicons and language models. Specific data for a more targeted study into lexical differences was provided by Keuleers et al. (2015). They conducted a large crowd-sourcing experiment, in which test subjects had to indicate for Dutch words and pseudowords whether or not they recognized the presented forms as Dutch words. On the basis of the collected data, they could study the influence of factors like age, education level and proficiency in other languages on vocabulary size. However, they also compiled a table which reports for each word which percentage of participants in Belgium and the Netherlands recognized the word, something they called *prevalence*. They pose that prevalence is complementary to corpus-based word frequency counts for the prediction of word occurrence. For more rare words, prevalence should be better, as these words will likely be absent, or show very low counts, in corpora. They prove their point by using both prevalence and frequency data (from SUBTLEX-NL; Keuleers et al, 2010) to predict reaction times from a lexical decision task. For the words for which both measures are present, prevalence and frequency have only a correlation of 0.35, showing they are really different. Log frequency predicted 36% of the variance in the reaction times, prevalence 33%, and jointly they predicted 51%. Whereas their analyses were conducted on the full set of measurements, the prevalence table contains separate values for Belgian and Netherlandic Dutch, which we will use below.

Studies investigating syntactic differences between Belgian and Netherlandic Dutch generally focus on specific constructions. The differences that are observed seldom concern constructions that are unique to either language variety. Mostly constructions are found to occur in both varieties with a preference for one construction over another, often under specific conditions and in specific contexts. This requires intricate analyses to bring to light the complexes of syntactic and semantic/pragmatic factors that can explain the subtle differences in the way the constructions are used. Examples are the studies by Grondelaers and Speelman (2007) on presentative sentences, Boogaart (2007) on conditional clauses with *moest(en)* and *mocht(en)*, Barbiers and Bennis (2010) on constituent ordering in the clause-final verb group and Gyselinck and Colleman (2016) on the intensifying use of the fake reflexive resultative construction.

3 Experimental Data

The data for the DFS shared task of VarDial2018 (Zampieri et al., 2018) originate from the SUBTIEL Corpus (van der Lee, 2017; van der Lee and van den Bosch, 2017). They consist of Dutch subtitles for movies and tv shows, produced by the Dutch and the Belgian branch of the company Broadcast Text

Feature type	Example	Total number	Number with odds ≥ 2
Char 1-gram	C1_;	70	2
Char 2-gram	C2_ZE	2,249	647
Char 3-gram	C3_op!	19,868	5,631
Char 4-gram	C4_DiMe	90,880	27,497
Char 5-gram	C5_Sami#	242,637	77,472
Token 1-gram	T1_W_Text	47,272	19,222
Token 2-gram	T2_WW_#_Oke	202,681	78,682
Token 3-gram	T3_WWW_de_dingen_des	264,367	107,595

Table 1: Lexical features

International (BTI Studios). For the shared task, the subtitles have been marked as either Belgian or Netherlandic Dutch depending on the market for which they were prepared. This is likely, but not guaranteed, to correspond to subtitling in the corresponding branch by a native speaker of the corresponding variety.

Sequences of subtitles comprising a number of full subtitles and containing about 30-35 words were selected randomly from the whole data set to be used as task items. 150,000 training items and 250 development items for each variety were provided beforehand, and 20,000 test items a few days before the submission deadline. As we did not use a tuning step in our training, we merged the development data with the training data. The training and test data were both selected completely randomly from the full data set. There was no overlap in items, but it was possible that there were test items belonging to the same movies or tv shows as training items. As we will see below, this had a substantial influence on the nature of the task.

When inspecting the data, we found several artefacts of earlier preprocessing steps. Most notably, all characters with diacritics were removed (e.g. *één* ("a") became *n*), or alternatively the diacritic was removed but also a space was inserted (e.g. *ruïne* ("ruin") became *ru ine)*. Furthermore, periods in numbers had spaces inserted next to them (e.g. *20.000* ("20,000") became *20. 000*). Also apostrophes in words like *z'n* (*zijn*, "his") were removed; apparently, some correction had already been applied, but this also produced non-existing forms like *zeen*.

Now, these artefacts would not be a problem for character or token n-gram recognition. However, we were planning to use POS tagging and syntactic parsing, for which these artefacts would most certainly lead to errors. We therefore decided to include a preprocessing step in which we tried to correct most of these artefacts. For the diacritics, it would have been easiest to compare to a Dutch word list in which diacritics were included, but we did not manage to acquire such a resource quickly enough. Instead, we inspected derived word counts and the text itself manually, and build a list of about 270 regular expression substitutes, such as

```
  s/\([Gg]e\) dealiseerde /\1idealiseerde /g
  s/i re /iere /g
  s/\([0-9]\)\([,.]\)  *\([0-9]\)/\1\2\3/g
```

As we spent only limited time on this, we missed cases, even (in retrospect) obvious ones like *financi ele* (*financiële*, "financial"), leading to non-words like *ele* in the observations below.

4 Recognition Features

As stated above, our main goal was to find differences between the Northern and Southern varieties of standard Dutch, both lexical and syntactic. To be able to extract features needed for this goal, we analysed the text with a combination of Frog (van den Bosch et al., 2007) and Alpino (Bouma et al., 2001). However, we also took the recognition task seriously, and included more traditional character and token n-gram features (Stamatatos, 2009). In the actual recognition we only used features with odds higher or equal to 2 in favour of either variety (see Section 5).

The character and token n-gram features (below called *lexical features*) were extracted from the original (but cleaned up) data. Hash characters (#) were inserted before and after each sentence for the extraction of begin/end n-grams. For character n-grams, n ranged from 1 to 5, for token n-grams from 1 to 3. Table 1 shows some examples and statistics for the lexical features.

```
Voor        VZ(init)
vandaag     BW()
had         WW(pv,verl,ev)
ik          VNW(pers,pron,nomin,vol,1,ev)
al          BW()
een         LID(onbep,stan,agr)
maand       N(soort,ev,basis,zijd,stan)
geen        VNW(onbep,det,stan,prenom,zonder,agr)
stem        N(soort,ev,basis,zijd,stan)
meer        VNW(onbep,grad,stan,vrij,zonder,comp)
gekregen    WW(vd,vrij,zonder)
.           LET()
```

Figure 1: POS tagging example.

Feature type	**Example**	**Total number**	**Number with odds ≥ 2**
Tagging 1-gram	`T1_P_WW(inf,prenom,zonder)`	34,771	14,527
Tagging 2-gram	`T2_GL_LID(bep)_redder`	1,320,282	497,488
Tagging 3-gram	`T3_GWG_TW(hoofd)_a_TW(hoofd)`	11,640,773	4,379,149

Table 2: Tagging features

The next level of features are those which can be extracted after POS tagging (below called *tagging features*). Frog yields an annotation with the tagset created for the Spoken Dutch Corpus (Oostdijk et al., 2000). An example is shown in Figure 1. We used the POS tags by themselves, but also in broader POS groups, retaining only the first attribute, leading to groups like *WW(inf)* and *N(soort)*. Also, apart from the POS tags, the tagging process provides us with lemmas, or rather stems as e.g. for *was* ("was"), we find the first person singular *ben* ("am"). From this annotation we derived unigrams, bigrams and trigrams. In each of the positions of the n-grams, we put one of the following: the word (`W`), the lemma (`L`), the full POS tag (`P`), or the POS group (`G`). As an example, `T3_GLP_LID(bep)_ding_LID(bep,gen,evmo)` is the trigram built with a POS group, a lemma and a POS, that corresponds to the example *de dingen des* ("the things of") in Table 1. Table 2 shows some examples and statistics for the tagging features. Here the pure word combinations are excluded as they have been counted as lexical features.

The final group of features (below called *syntactic features*) have been derived from the syntactic parse produced by Frog and Alpino. However, since the dependency structure is less amenable to variation studies than a constituency structure, we first transformed the trees. We started with the 'surfacing' procedure developed by Erwin Komen (2015), and followed it up with a few more transformations, especially around the verb phrase. Furthermore, the analyses were lexicalized by percolating the head words upwards. As an example the parse of the sentence in Figure 1 is shown in Figure 2. From the trees we derived two types of features. We built a kind of syntactic n-grams by taking subtrees such as a functional constituent (F) realized by syntactic category (C) containing a functional constituent realized by syntactic category (e.g. `SFCFC_mod_WHREL_obj1_TW`, a modifier realized by a *wh*-relative clause (`WHREL`) containing a direct object (`obj1`) realized by a cardinal numeral (`TW`); or `SCFFCCL_NP_hd_N(ding)_mod_NP(leven)`, a noun phrase (`NP`) containing both a head (`hd`) realized by a noun (`N`) with lemma *ding* and a modifier (`mod`) realized by a noun phrase (`NP`) with a head *leven*, which corresponds to the example *de dingen des levens* ("the things of life") which we already saw above. The second type of feature are the full rewrites at all positions in the tree, e.g. `SRFC_WHQ_whd_VNW_hd_WWlex_obj1_TW_<NOFUN>_.` (a *wh-* question realized by a interrogative pronoun, a verb, and a direct object realized by a cardinal numeral, ending with a sentence closing

```
<NOFUN>:SMAIN(krijg) ->      [ mod auxv su mod obj1 mod lexv <NOFUN> ]
      mod:PP(voor|vandaag) ->      [ hd obj1 ]
          hd:VZ(voor) -> voor
          obj1:BW(vandaag) -> vandaag
      auxv:WWaux(heb) -> heb
      su:VNW(ik) -> ik
      mod:NP(maand) ->      [ mod det hd ]
          mod:BW(al) -> al
          det:LID(een) -> een
          hd:N(maand) -> maand
      obj1:NP(stem) ->      [ det hd ]
          det:VNW(geen) -> geen
          hd:N(stem) -> stem
      mod:VNW(meer) -> meer
      lexv:WWlex(krijg) -> krijg
      <NOFUN>:.(.) -> .
```

Figure 2: Syntactic analysis example

Feature type	**Example**	**Total number**	**Number odds ≥ 2**
Subtree (not lexicalized)	`SCFCFC_NP_mod_REL_predc_PPRES`	189,411	46,853

Subtree (lexicalized)	`SCFFCCL_CP_cmp_VG(als)_`
	`lexv_WWlex(overlijd)`
		630,926	242,976

Rewrite (to functions only)	`SRF_SMAIN_su_auxv_lexv_predc`	12,746	3,594
Rewrite to functions and categories	`SRFC_NP_det_LID_hd_N_` `mod_PP_mod_PP_mod_PP`	36,545	11,756

Table 3: Syntactic features

punctuation mark (even the question mark has the category label .). Table 3 shows some examples and statistics for the syntactic features.

5 Recognition System and Results

Our choice of recognition system as well was influenced by our goal of finding differences between the two language varieties. After successful recognition, we wanted to be able to identify which features contributed to the success. For this experiment, we chose a very simple algorithm. We counted the occurrences of each feature in the Netherlandic and Belgian training items and compared the two counts to derive odds. For example, the word bigram *Komaan_,* ("Come on ,") was found 209 times in the Belgian items and 4 times in the Netherlandic items, leading to odds of 52.25 in favour of Belgian Dutch.

If a feature was not seen in one of the varieties, its count was set to 1 for the calculation of odds, e.g. the name *Sami* was found 275 times in the Belgian items and never in the Netherlandic items, leading to odds of 275. For the actual recognition, we only used features that had odds higher than or equal to 2 in either direction. In the numerical representations below, odds in favour of Netherlandic are shown as negative and in favour of Belgian as positive.

In the test phase, all features present in an item were taken and their odds contributed directly to the item score. In the simplest version, all odds were simply added, after which a positive total indicated

Method/features	Accuracy	F1micro	F1macro	F1weighted
Simple/all	0.6406	0.6406	0.6403	0.6403
Voted/lexical	0.6344	0.6344	0.6342	0.6342
Voted/tagging	0.6343	0.6343	0.6341	0.6341
Voted/syntactic	0.6142	0.6142	0.6134	0.6134
Voted/lexical+tagging+syntactic	0.6458	0.6458	0.6456	0.6456

Table 4: VarDial2018-DFS scores for various approaches

Word (Netherlandic)	Training data odds	#correct/ wrong in test		Word (Belgian)	Training data odds	#correct/ wrong in test
!	-13.19	59/4		da’s	8.83	68/5
EEN	-264.00	22/0		Hope	4.52	30/1
MUZIEK	-482.00	26/2		Shawn	9.28	28/2
Oke	-228.00	19/0		Sami	289.00	21/0
gerecht	-2.750	48/11		Bo	6.51	24/1
MasterChef	-55.00	14/0		Lucas	3.72	22/1
inmiddels	-2.66	18/2		komaan	49.71	21/1
namelijk	-2.030	20/3		amuseren	5.51	24/2
melding	-2.68	11/0		plots	8.97	19/1
Foreman	-18.00	11/0		aanvaarden	3.16	19/1

Table 5: Top distinctive words for test set

Belgian and a negative total Netherlandic. Based on experience in other projects, we expected an increase in recognition quality when taking several feature classes and then combining the results. For such an approach, we split the features split into 15 classes. The lexical features `C1, C2, C3, C4, C5, T1_W, T2_WW` and `T3_WWW` each formed their own class. N-grams built purely from POS tags formed classes `T1_tag, T2_tag` and `T3_tag`. N-grams built with other component mixes formed classes `T2_mix` and `T3_mix`. Syntactic features were split into `Spure` and `Slexical`, depending on whether they contained references to lexical items. The odds addition was done per feature class, leading to a vote for Belgian or Netherlandic, or no vote if no features of the class were present. The 15 votes were combined without weighting and the variety with most votes was selected as final result. If the varieties had an equal number of votes, Belgian was selected, a heuristic based on some experiments within the training data.

The results of various settings are shown in Table 4. Looking at the processing needed for specific features, and taking the overall groups lexical features, tagging features and syntactic features, we see that all three groups perform worse than the simple odds addition of all features together. Voting with all features, however, outperforms the simple addition. Syntactic features by themselves perform quite poorly, and would have ended up at rank 5 in the shared task; lexical and tagging features by themselves at rank 3. However, their combination and the addition of the syntactic features pushes the result significantly higher, to rank 2, demonstrating the value of both system combination (widely accepted) and syntactic information (not widely accepted). Still, despite access to more informative features, we must admit defeat to team Tübingen-Oslo (Çöltekin et al., 2018), who reach a score of 0.6600 using an SVM classifier based on character and word n-grams. This we attribute to our choice of recognition method, which as already mentioned was based on explanatory power more than recognition power.

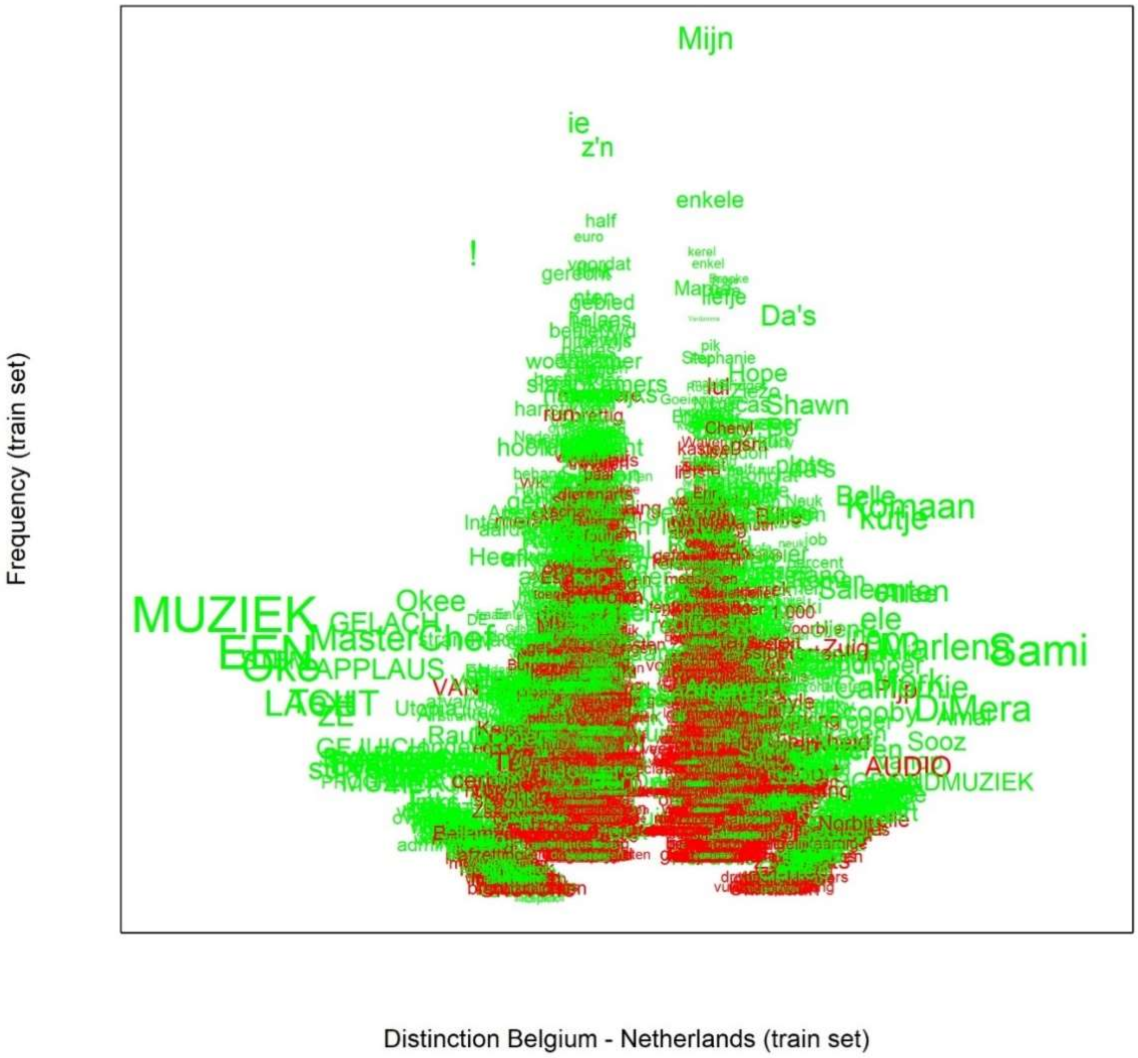

Figure 3: Word bias in training and test set. The horizontal position represents bias, with bias towards Netherlandic on the left and bias towards Belgian on the right. The vertical position represents the frequency in the training data (on a log scale). The colour represents whether the bias was the same in the test data (green) or not (red). Finally, the size represents how useful the feature proved in judging the test set (calculated as (#correct – 3*#incorrect) * odds).

6 Distinction Power of Individual Tokens

Seeing that the Southern and Northern varieties of Dutch can apparently be distinguished to some degree, we would like to know which features contribute to this distinction, in other words what the actual differences between the varieties are. In this section, we focus on individual tokens. Not only is this a stated focus of the VarDial workshop, but we can also compare the results with our intuitions. The observations for tokens can later on help in the examination of syntactic features (Section 7).

Table 5 shows the most useful features for the test set (here correlated features are ignored) and Figure 3 visually represents the usefulness of all features applied for the test set.[1] In both we see that it is

[1] The sources files for Figures 3 and 4, as well as the underlying data, can be found at https://cls.ru.nl/staff/hvhalteren/VarDial2018_DFS_Taurus_Support.zip

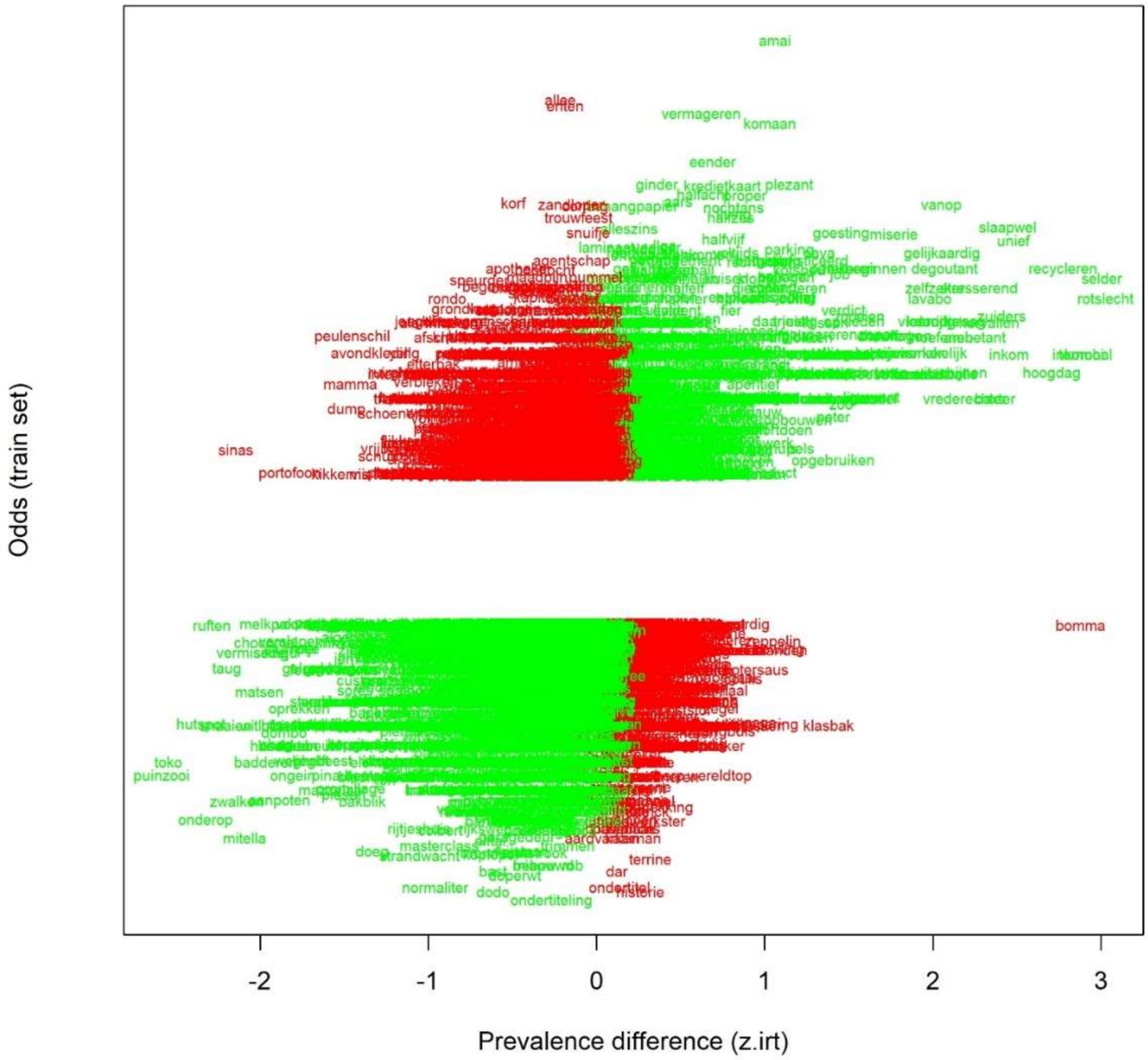

Figure 4: DFS bias compared to prevalence (Keuleers at al., 2015) for words present in both data sets. The horizontal position represents prevalence (z-scores) and the vertical positions odds in the DFS training data (log scale). The colour again indicates whether the two measurements agree on assigning the corresponding language variety.

certainly not only differences in language variety that we are measuring. First of all, we see many proper names, with as most striking example *Sami*, not the Finnish people, but a character from the soap series Days of our Lives. She occurs 289 times in the training set and 21 times in the test set, always on the Belgian side. On the Netherlandic side, we find the tv show *MasterChef*, complemented by the word *gerecht* ("dish"). Other words might also be linked to specific content: a remarkable number of police-connected words, such as *melding* ("report"), *bekeuring* ("traffic ticket") and *bestuurder* ("driver") occur on the Netherlandic side. Next we have a cluster of all-capital words, such as *MUZIEK* ("MUSIC"). These appear to be descriptions of background noises. They occur mostly on the Netherlandic side, with a few exceptions, e.g. *ACHTERGRONDMUZIEK* ("BACKGROUND MUSIC") on the Belgian side. This too is not related to language variety but more likely to the nature of subtitling in the two branches of BTI. The same might be true for the exclamation mark, which together with the semicolon are the only C1 features with odds greater than 2 (13.19 and 18.00).

Still, we also observe words which do seem linked to a specific language variety. Interjections like *komaan* ("come on") and *Oke* ("Right") also intuitively belong to Belgian and Netherlandic Dutch, respectively. The same can be said for *da's* ("that's"; Belgian, but probably extending into the South of The Netherlands), *ie* ("he"; Netherlandic) and *plots* ("suddenly"; Belgian). For the other visible words, we do not have clear intuitions. Most of them we will therefore check against another resource in the next paragraph. Before that, we want to mention the capitalized *Mijn* ("my"), which might have a stylistic rather than a lexical cause.

As mentioned in Section 2, Keuleers et al. (2015) provided prevalence measurements for (about 54,000) Dutch words (downloadable through http://crr.ugent.be/archives/1796). Figure 4 shows the relation between the difference in prevalence in Belgium and The Netherlands (using the irt z-scores also preferred by Keuleers et al.; for an explanation, see there) and the odds from our own training data. We include all lemmas which a) occur in the prevalence table and b) have odds higher than 2 in our measurements. The visual impression that the two resources agree more often than they disagree is correct: they agree 3,717 times (2,628 Netherlandic and 1,089 Belgian) while they disagree 2,060 times. Still, they disagree often enough to conclude that, if we accept prevalence as a good indicator of language variety bias, the DFS data is inadequate for proper identification of that bias. Furthermore, even high odds do not guarantee proper attribution. On the other hand, highly prevalent words do appear to be recognized also with the DFS data. Although we did not measure the exact same thing (difference versus sum), these observations are consistent with the findings by Keuleers et al. (2015) about the complementary nature of prevalence and corpus-based frequency counts.

7 Distinction Power of Syntactic Features

Our special interest in this experiment lay in syntactic differences between Northern and Southern Dutch. Ideally any different uses of syntax would be found by way of highly distinguishing syntactic features. However, our observations for the lexical features made us less optimistic. Trying at least to avoid syntactic features that were shadows of lexical ones, we filtered the set of syntactic features that we would examine: we used only rewrite features appearing in test items where the lexical features led to a wrong attribution but the tagging and syntactic features attributed correctly, and from the resulting feature list we then removed all syntactic features that were correlated to any lexical features. Upon manual inspection of the selected features we found various syntactic constructions that at least in these data point to Belgian or Netherlandic authorship. They give us a first handle as regards the potential syntactic differences that exist between the two varieties. Further research is needed to establish whether these are not just an artefact of the data. Below we present and discuss some of the constructions that were identified.

The use of constructions with anticipatory *het* were found to be associated with Belgian authorship. An example is *En het is vanwege jou dat ze deze sessies hebben georganiseerd*. ("And it is because of you that they have organised these sessions."). What is striking in the instances that we come across is that in all of them the subject complement (here: *vanwege jou*) is an adverb (*wel*, *niet*, *dus*) or, as in the example, a prepositional phrase.

Coordinations in which *dus* ("so") appears in the function of coordinator joining main clauses were associated with Netherlandic authorship. An example is *er gebeurde niks dus is ze dood.* ("nothing happened so now she is dead."). All cases share the same structure where a cause, circumstance or reason is given in the first conjoin of the coordination and the second conjoin introduced by *dus* relates the consequence(s), result, or the thing(s) that happened next.

As an apparent feature of Netherlandic Dutch we found that the final conjoin in a coordination is a word or phrase equivalent to *etcetera*. For Dutch we came across *enzovoort*, *et cetera*, but also *noem maar op*, as in for example *hij weet wel alles over de vissen in de zee over bomen en planten, noem maar op.* ("he knows everything about the fish in the sea about trees and plants, and what have you").

Then there are adverbial clauses in sentence-initial position in monotransitive declarative sentences. In standard Dutch in regular declarative sentences with unmarked word order, the subject precedes the verb operator. However, when an adverbial occurs sentence-initially, subject and verb operator are inverted. For Netherlandic Dutch the occurrence of monotransitive sentences with sentence-initial adverbial clauses was found to be a distinctive feature. These initial adverbial clauses included both conditional clauses (typically introduced by *als* ("if") as well as temporal adverbial clauses (for example introduced by *nadat* ("after"), *terwijl* ('while"), or *toen* ("when"), as in *Toen we gingen dansen speelden ze dit liedje.* ("When we went dancing they played this song.").

Another feature associated with Netherlandic Dutch was the use of *wh*-clauses as direct object. For example, *U weet dus wat er in zit.* ("You know what is in there.").

Finally, one more feature typically associated with Belgian Dutch was the use of an imperative clause followed by a form of address as for example in *Geef hier dat geld, vuile hufter.* ("Hand over that money, you bastard.").

8 Conclusion

We built a recognition system to distinguish between Belgian and Netherlandic Dutch subtitles, as provided for the DFS shared task at VarDial2018. We used a wide range of features, spanning from character n-grams to syntactic rewrites. As our primary goal was to identify features which differed between the two language varieties, we used a simple marker-based recognition method, with which we could measure directly how informative each feature was for the test data.

As for the participation in the shared task, we can judge our results to be positive. Achieving the second place in the ranking shows that even a simple marker-based method can hold its own in a competition which we expected to be dominated by more intricate machine learning methods. We can only assume that our wide range of knowledge-inspired features, including fully syntactic ones, made up for the weaker method.

However, as for identifying differences between Belgian and Netherlandic Dutch, we have to view our current results as merely a first step. We did manage to identify some features that appear to be biased towards either Belgian or Netherlandic Dutch, but it is as yet unclear if this is because of the language variety of the source text. There were too many interfering factors to be sure, such as topic (the movie or show the subtitles were from), differing genres (subtitling conventions in the local branches of BTI), and processing difficulties (not quite optimally appropriate software, run on not quite clean text). Still, we did manage to identify some potential syntactic differences between Belgian and Netherlandic Dutch, our main goal, while taking precautions to avoid interference from these factors. As a result, we do have a basis for research on further data, in which we can try to confirm our findings.

Acknowledgements

We thank Erwin Komen and Micha Hulsbosch for preparing a script for the analysis of the text with Frog, Alpino and the surfacing software.

Reference

R. Harald Baayen, Hans van Halteren, and Fiona Tweedie. 1996. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3), 121‑131.

Sjef Barbiers and Hans Bennis. 2010. De plaats van het werkwoord in zuid en noord. In: Barbiers, S., H. Bennis, J. De Caluwe & J. Van Keymeulen (red.). *Voor Magda. Artikelen voor Magda Devos bij haar afscheid van de Universiteit Gent*. Gent: Academia Press, 25-42.

Ronny Boogaart. 2007. Conditionele constructies met *moest(en)* en *mocht(en)* in Belgisch-Nederlands en Nederlands-Nederlands. *Neerlandistiek.nl*. 07.05

Gosse Bouma, Gertjan Van Noord, and Robert Malouf. 2001. Alpino: Wide-coverage computational analysis of Dutch. *Language and Computers,* 37 (2001): 45-59.

Çağrı Çöltekin, Taraka Rama, and Verena Blaschke. 2018. Tübingen-Oslo Team at the VarDial 2018 Evaluation Campaign: An Analysis of N-gram Features in Language Variety Identification. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial), Santa Fe, USA.*

Julien Despres, Petr Fousek, Jean-Luc Gauvain, Sandrine Gay, Yvan Josse, Lori Lamel, and Abdel Messaoudi. 2009. Modeling Northern and Southern varieties of Dutch for STT. *Tenth Annual Conference of the International Speech Communication Association.*

Stefan Grondelaers and Dirk. Speelman. 2007. A Variationist Account of Constituent Ordering in Presentative Sentences in Belgian Dutch. *Corpus Linguistics and Linguistic Theory*, 3(2): 161-193.

Emmeline Gyselinck and Timothy Colleman. 2016. Je dood vervelen of je te pletter amuseren? Het intensiverende gebruik van de pseudoreflexieve resultatiefconstructie in hedendaags Belgisch en Nederlands Nederlands. HANDELINGEN: KONINKLIJKE ZUID-NEDERLANDSE MAATSCHAPPIJ VOOR TAAL-EN LETTERKUNDE EN GESCHIEDENIS 69 (2016): 103-136.

Emmanuel Keuleers, Kevin Diependaele, and Marc Brysbaert. 2010. Practice Effects in Large-Scale Visual Word Recognition Studies: A Lexical Decision Study on 14,000 Dutch Mono- and Disyllabic Words and Nonwords. *Frontiers in Psychology*, 1.

Emmanual Keuleers, Michaël Stevens, Paweł Mandera, and Marc Brysbaert. 2015. Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *Quarterly Journal of Experimental Psychology*, 68, 1665-1692.

Erwin Komen. 2015. Surfacing Dutch syntactic parses. Presentation at Computational Linguistics in the Netherlands (CLIN26), Amsterdam, 2015. http://wordpress.let.vupr.nl/clin26/abstracts/

Nelleke Oostdijk. 2000. The Spoken Dutch Corpus: Overview and first evaluation, *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*, 887–894.

Nelleke Oostdijk, Martin Reynaert, Veronique Hoste, and Ineke Schuurman. 2013. The construction of a 500-million-word reference corpus of contemporary written Dutch. In *Essential speech and language technology for Dutch*. Springer, Berlin, Heidelberg. 219-247.

Efstathios Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology.* Volume 60 Issue 3, March 2009. 538-556.

Antal van den Bosch, Bertjan Busser, Walter Daelemans, and Sander Canisius. 2007. An efficient memory-based morphosyntactic tagger and parser for Dutch, In F. van Eynde, P. Dirix, I. Schuurman, and V. Vandeghinste (Eds.), *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting*, Leuven, Belgium, 99-114.

Chris van der Lee. 2017. *Text-based video genre classification using multiple feature categories and categorization methods.* Master's Thesis. Tilburg University.

Chris van der Lee and Antal van den Bosch. 2017. Exploring Lexical and Syntactic Features for Language Variety Identification. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 190–199, Valencia, Spain.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial Evaluation Campaign 2017. *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*. Valencia, Spain.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Ahmed Ali, Suwon Shon, James Glass, Yves Scherrer, Tanja Samardžić, Nikola Ljubešić, Jörg Tiedemann, Chris van der Lee, Stefan Grondelaers, Nelleke Oostdijk, Antal van den Bosch, Ritesh Kumar, Bornini Lahiri and Mayank Jain. 2018. Language Identification and Morphosyntactic Tagging: The Second VarDial Evaluation Campaign. *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial).* Santa Fe, USA.