Modeling Syntactic Dependencies in Southern Dutch Dialects

Loic De Langhe, Jasper Degraeuwe, Melissa Farasyn, Véronique Hoste
Ghent University, Belgium
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
firstname.lastname@ugent.be

Abstract

Dependency parsing of non-normative language varieties remains a challenge for modern NLP. While contemporary parsers excel at standardized languages, dialectal variation – for instance in function words, conjunctives, and verb clustering – introduces syntactic ambiguity that disrupts traditional parsing approaches. In this paper, we conduct a quantitative evaluation of syntactic dependencies in Southern Dutch dialects, leveraging a standardized dialect corpus to isolate syntactic effects from lexical variation. Using a neural biaffine dependency parser with various mono- and multilingual transformer-based encoders, we benchmark parsing performance on standard Dutch, dialectal data, and mixed training sets. Our results demonstrate that incorporating dialect-specific data significantly enhances parsing accuracy, yet certain syntactic structures remain difficult to resolve, even with dedicated adaptation. These findings highlight the need for more nuanced parsing strategies and improved syntactic modeling for non-normative language varieties.

1 Introduction

Modern language models demonstrate impressive mastery of both semantics and syntax across most well-studied languages. This success is largely due to abundant training data or effective transfer learning methods (Weiss et al., 2016), which ensure strong alignment for mid-to-high-resource languages. However, processing non-normative variants such as dialects and code-switched speech remains a significant challenge (Jørgensen et al., 2015). In this paper, we present a comprehensive evaluation and analysis of syntactic dependencies in various Southern Dutch language varieties based on a sizable human-annotated corpus of transcribed and to a certain extent standardized dialect speech (Breitbarth et al., 2020; Ghyselen et al., 2020; Breitbarth et al., 2024).

The Southern Dutch dialect (SDD) group encompasses dialects spoken in (i) Dutch-speaking Belgium, (ii) the three southern provinces of the Netherlands (Limburg, Noord-Brabant and Zeeland) and (iii) the Flemish-speaking dialect region in France (Farasyn et al., 2022). A geographical situation of the area, including the main dialect variants, can be found in Figure 1. SDDs can be grouped into four major varieties. Flemish includes West Flemish, East Flemish, Zeeland Flemish (spoken in Zeeland Flanders, i.e. south of the Westerschelde in the province of Zeeland, excluding the Land van Hulst), and the nearly extinct French Flemish (spoken in northern France). Zeelandic is spoken in the other areas of Zeeland. Brabantic covers North Brabant, Antwerp, and Flemish Brabant, while Limburgish is spoken in Belgian and Dutch Limburg.

In this study, the term dialect refers specifically to historically established regional varieties, distinct from *tussentaal* (‘interlanguage’). *Tussentaal* is a linguistic phenomenon where regional speech varieties gradually converge toward the standard language, leading to dialect erosion. Unlike traditional dialects, which have well-defined grammatical, phonological, and lexical features, *tussentaal* blends regional and standard elements (De Caluwe, 2009). While dialect knowledge in the Low Countries has steadily declined since the 1980s, *tussentaal* has not. Recent research shows that *tussentaal* itself displays both regional and social variation, functioning as a stratified cluster of varieties rather than a single intermediate form. At the same time, certain features show signs of supraregional stabilization, particularly in informal spoken registers, with both diversification and convergence depending on social context and speaker group (De Caluwe et al., 2013; Ghyselen, 2015).

Despite the decline in dialect use over the past decades, the linguistic diversity within the original dialect varieties remains a rich and interesting area

Proceedings of the Eighth International Conference on Dependency Linguistics (Depling, SyntaxFest 2025), pages 24–35
August 27-28, 2025 ©2025 Association for Computational Linguistics

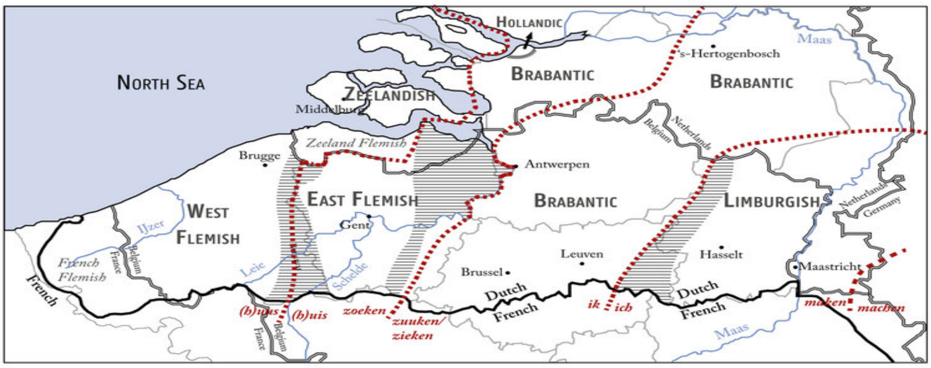

Figure 1: Geographical distribution of spoken dialects in the region, highlighting linguistic boundaries and areas of dialect convergence (Farasyn et al., 2022)

of study, both from a historical and computational perspective. The main variants exhibit both shared and individual syntactic particularities, which are not found in the standard Dutch language (Barbiers et al., 2005).

This paper is structured as follows: we first examine key syntactic features of SDDs in Section 2. Next, we benchmark a neural dependency parser with various mono- and multilingual transformer encoders (Section 3). We evaluate models trained on standard Dutch, dialect data, and both combined, focusing on four dialects: West Flemish, East Flemish, Brabantian, and Zeeland Flemish. Our results show that incorporating dialect data significantly improves parser performance, though some syntactic patterns remain challenging, and certain dialects pose difficulties even with dedicated training data (Section 4).

2 Related Work

2.1 Dependency Parsing

Dependency parsing has always been a popular research topic in the Dutch language domain. Early formal approaches struggled with cross-serial dependencies, illustrating the limitations of context-free grammars (Bresnan et al., 1987). The development of the Alpino Dependency Treebank (Van der Beek et al., 2002; Van Noord, 2006) provided a crucial resource that played a key role in the advancement of rule-based and statistical parsers, driving further progress in the field.

The rise of Universal Dependencies (UD) (Nivre et al., 2016) standardized Dutch dependency annotation, fostering cross-linguistic research and improving multilingual parsing. Dutch UD treebanks, such as LassySmall UD and Alpino UD (Bouma and van Noord, 2017), have supported data-driven models, including transition-based (Nivre et al., 2006) and graph-based (McDonald et al., 2006) approaches. More recently, biaffine dependency parsing (Dozat and Manning, 2016) combined with transformer encoders such as BERTje (De Vries et al., 2019) and RobBERT (Delobelle et al., 2020) have achieved state-of-the-art results for Dutch. Additionally, more advanced techniques such as self-distillation (de Kok and Pütz, 2020) have also been proposed as methods for further enhancing modern-day neural parsers.

Dependency parsing of non-normative language, such as dialects and code-switching, is challenging due to linguistic variability and scarce annotated data (Jørgensen et al., 2015). While some methods use domain adaptation and transformer models to improve robustness (Jørgensen et al., 2016; Nguyen et al., 2020), processing dialect remains a largely open problem. For Dutch specifically, UD-based dialect treebanks (Braggaar and van der Goot, 2021) and transfer learning have enhanced parsing for northern regional varieties and informal registers (Braggaar and Van Der Goot, 2021). While some progress has been made with these language variants, the limited availability of data remains a significant obstacle. In our own work on SDDs, we aim to tackle this challenge by incorporating the substantial (and partly standardized) GCND corpus (Breitbarth et al., 2024), which helps mitigate two major concerns that commonly affect studies on non-normative language: spelling variation and limited data availability. The corpus includes two transcription layers, both of which use Dutch-based orthographic conventions to reduce phonological variation. The first layer remains dialectal in morphology, syntax, and vocabulary, while the second 'dutchified' layer adds light morphological and lexical normalization to facilitate readability and automatic processing (Ghyselen et al., 2020). All parsing experiments in this paper were conducted on the second transcription layer.

2.2 Syntactic Variations in Southern Dutch Dialects

SDDs exhibit considerable syntactic variation (Barbiers et al., 2005), posing challenges for traditional parsers. These dialects differ from standard Dutch in areas such as word order, the usage and placement of function words, pronominal paradigms, negation, complementizer systems and regionally specific conjunctives, all of which result in widely varying parse trees. Additionally, clitic doubling and verb cluster variation further challenge parsing, as standard models struggle to map these structures onto expected patterns. In the paragraphs below, we address six linguistic phenomena that are likely to significantly hinder the performance of a parser trained on standard language.

Subject doubling (or tripling) is a phenomenon where the subject of a sentence occurs multiple times, typically involving a combination of a pronoun and a full noun phrase or two pronouns (De Vogelaer, Gunther, 2006). This construction is particularly common in the SDDs, including West Flemish, East Flemish, Brabantic, and Limburgish varieties (Van Craenenbroeck and Van Koppen, 2002). Subject doubling often occurs as part of topic marking or emphasis, distinguishing these dialects from standard Dutch, where such constructions are ungrammatical.

1. **Ze** werkt **zij** in Brussel

   EN: She works (she) in Brussels

Here, the first subject (*Ze*) serves as a topic, while the second subject (*zij*) functions as a resumptive pronoun. In some dialects, particularly in West Flemish and Brabantian, subject doubling can extend even further to subject tripling, where a noun phrase is followed by two pronouns (De Vogelaer and Devos, 2008).

Negation doubling and tripling is another syntactic feature observed in various southern dialect varieties (Haegeman and Zanuttini, 1996). This phenomenon involves the repetition of negation markers within a single sentence, often used for emphasis or to express stronger negation.

2. Ik heb dat **nooit niet** gedaan.

   EN: I never (not) did that

Here, the negation particle *niet* (not) is doubled with the negative indefinit *nooit* (never), which intensifies the negation beyond what is found in standard Dutch. In some dialects, negation can be tripled, further emphasizing its intensity. The phenomenon of negation stacking in dialects has been thoroughly discussed in the literature (Paardekooper, 2015; De Schutter, 2015) and contrasts with the grammatical use where two negative elements typically cancel each other out, creating a positive meaning.

***En* negation** In many Southern Dutch dialects, negation can involve a preverbal particle *en*, in addition to a negation particle like *niet* or a negative indefinite like *nooit*. This results in negation doubling or tripling, depending on the combination. For instance, a sentence such as:

3. Ik **en** zie niets.

   EN: I (and) don't see anything

The *en* particle is a remnant of the historical sentential negator, now reinterpreted as a discourse-related element (Breitbarth and Haegeman, 2014, 2015). Its surface form coincides with the standard coordinating conjunction *en* purely by chance since it has a distinct syntactic origin and function. The construction remains productive in many Southern Dutch dialects, particularly in Flemish varieties (Neuckermans, 2008; Barbiers et al., 2007). It poses a challenge for automatic parsing due to its surface ambiguity and non-standard word order.

The expletive *dat* (that) can serve as an expletive in spoken dialect Dutch, usually following interrogative pronouns, relative pronouns or subordinating conjunctions, resulting in so-called complementizer doubling, as it overlaps in form with the standard complementizer (Bacskai-Atkari, 2020; Barbiers, 2009). While this complementizer does not change sentence meaning, it does change the word order compared to standard Dutch.

5. Ik weet niet waar **dat** hij is.

   EN: I don't know where (that) he is

This linguistic feature occurs frequently across all Flemish regions, with the notable exception of Southeast Limburg (Barbiers et al., 2007; Taeldeman, 2008).

Distinct comparative conjunctions occur widely in the SDDs (Rooy, 1965). Instead of the standard Dutch *dan* ('than'), many varieties use *als* ('as') or *of* ('or') in comparative constructions, as in:

4. Hij is groter **als** jou.

   EN: He is bigger than you

This variation is well attested across dialect areas (see also (Postma, 2006)). The forms are fully grammatical in their dialects, but are likely to confuse parsers trained on standard Dutch.

Deviating clause introductions are a final phenomenon commonly observed in dialects and informal speech. In standard Dutch, non-finite clauses with a to-infinitive are introduced either by the complementizer *om* or by a null element. *Om* is mandatory in conditional clauses but optional when the clause functions as a true subject, direct object, or postmodifier of a noun (Vandeweghe, 1971). In non-standard Flemish registers, *voor* and *van* are often used as an alternative to introduce non-finite clauses with a to-infinitive (Barbiers et al., 2005).

6. Ze deed dat **voor** beter te horen.

   EN: She did that to (for) hear better

3 Experiments

3.1 Data

3.1.1 Standard Dutch

The standard Dutch portion of our data consists of two benchmark corpora, annotated for syntactic dependencies. First, the Lassy-Small Corpus (Van Noord et al., 2013) contains a total of 1 million words sourced from the larger D-COI (50 million words) corpus (Oostdijk, 2006). The second part of our standard dataset consists of the Alpino treebank, a collection of over 150,000 words of newspaper data (Van der Beek et al., 2002). Unlike the Lassy Corpus, the Alpino treebank was synthetically created through the use of the eponymous Alpino parser, a HPSG-based linguistic analysis tool for parsing Dutch text (Van Noord, 2006).

For both corpora we use the official UD versions (Bouma and van Noord, 2017) which are made available through the Universal Dependencies project (Nivre et al., 2016).

3.1.2 Southern (Dialect) Dutch

Our dialect data is a subset of the larger Corpus of Spoken Dutch Dialects (GCND) (Breitbarth et al., 2020), which is part of the Voices of the Past project (Hellebaut et al., 2021). This data originates from a collection of audio-recorded interviews with native speakers, conducted over a 13-year period (1963–1976) (Hellebaut et al., 2021). In total, the project amassed over 700 hours of spoken dialect material from more than 500 distinct locations across Flanders and the Netherlands. In recent years, conservation efforts have ensured that the majority of this audio material has been transcribed and normalized for spelling (Ghyselen et al., 2020). As part of this project, a portion of the data has been annotated with POS tags and syntactic dependencies following the Alpino guidelines (Breitbarth et al., 2020). For the annotation of syntactic dependencies this process follows a two-step approach: first, transcriptions are processed using the Alpino syntactic parser, and then they undergo manual correction by human annotators (Farasyn et al., 2022). At this stage, the syntactic annotations adhere to the original Alpino annotation guidelines. To create a usable corpus in UD format, we follow the same process [1] as outlined in previous work (Bouma and van Noord, 2017).

We have access to a total of 26,146 sentences from four dialect (sub)groups: Zeeland Flemish (4,141 sentences), Brabantian (4,496 sentences), East Flemish (8,674 sentences), and West Flemish (8,687 sentences). In these sentences, the language has been standardized to minimize lexical interference, allowing us to focus on syntactic patterns specific to these dialects, such as the ones discussed earlier in Section 2.2. Since dialectal varieties do not exhibit identical syntactic patterns, it is useful to first provide a high-level overview of how these characteristics vary across the different dialects. Figure 2 presents the distribution with which each characteristic appears in each of the four dialects.

3.2 Methodology

We train several versions of Diaparser (Attardi et al., 2021), a transformer-based extension on the standard deep biaffine Dependency Parser by Dozat and Manning (2016). In this framework, the encoder, typically a BERT-based transformer model, generates contextualized token embeddings that are fed into a biaffine network. This network then pre-

[1] https://github.com/rug-compling/alud

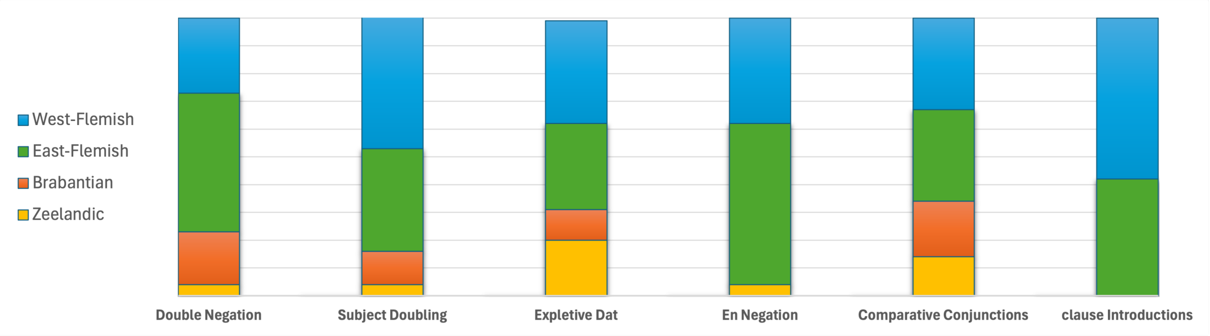

Figure 2: Syntactic pattern distribution across the four dialects.

Training Dataset	Encoder	UCM	LCM	UAS	LAS
Standard	*BERTje*	62.26	47.42	82.01	75.46
	RObBERT-2023	61.99	54.16	82.45	75.30
	ModernBERT	52.70	37.63	75.32	66.93
	mBERT	57.02	40.88	78.99	71.33
Dialect	*BERTje*	70.98	58.28	88.12	83.11
	RobBERT-2023	71.66	57.78	88.14	83.11
	ModernBERT	66.73	53.38	85.69	80.03
	mBERT	70.75	57.02	87.22	82.22
Merged	*BERTje*	74.07	62.91	89.07	85.17
	RobBERT-2023	73.77	60.61	89.26	84.86
	ModernBERT	71.59	57.48	87.84	82.84
	mBERT	72.89	59.66	88.46	83.85

Table 1: Main experimental results. Standard training data includes the Alpino and Lassy corpora; Dialect training data refers to the GCND corpus; the Merged dataset combines all three.

dicts the syntactic tree by simultaneously modeling both head and label prediction tasks. We select a variety of underlying encoder models to be tested on our data. These include monolingual models such the standard Dutch BERTje (De Vries et al., 2019) and the RoBERTa-based RobBERT-2023, a more recent Dutch language model (Delobelle et al., 2020). Additionally, we evaluate a benchmark multilingual encoder (mBERT) (Devlin et al., 2019) and the recently developed modernBERT (Warner et al., 2024).

For evaluation, we use a set of parsing metrics to assess model performance: Unlabeled Complete Match (UCM), Labeled Complete Match (LCM), Unlabeled Attachment Score (UAS) and Labeled Attachment Score (LAS). UCM and LCM are strict metrics that evaluate whether the entire predicted tree matches the gold-standard tree, with LCM requiring both the structure and labels to match exactly (Lücking et al., 2024). UAS measures the percentage of tokens with the correct head in the dependency tree, while LAS also considers the correct syntactic labels (Nivre et al., 2004).

3.3 Results

Table 1 presents a comprehensive overview of our main experimental results. The inclusion of dialect data leads to considerable performance gains. The best-performing model, BERTje, achieves an average improvement of 8.33% on the dialect dataset across all evaluated metrics, compared to a parser trained solely on standard (normative) Dutch. Furthermore, the best performance (mean improvement of 11.01%) is achieved when combining the both standard datasets (Alpino and Lassy) with dialect sentences, creating a well-balanced corpus that enhances the model's ability to process non-normative language fluently.

As expected, Dutch encoders outperform their multilingual counterparts, likely due to their specialized pretraining on Dutch linguistic structures. A key factor contributing to their strong performance is the absence of significant spelling variation, as the dialect datasets were normalized prior to training. While some dialect-specific vocabulary remains, tokenization issues appear to be minimal, especially compared to studies on less standardized dialects, where inconsistent orthography often hinders model performance (Jørgensen et al., 2015).

4 Discussion and Analysis

4.1 Performance and Syntactic Variation

While many Natural Language Processing studies on dialectal and non-normative data focus on lexical and spelling variation, our primary interest lies in how syntactic variation affects model performance. The SDDs discussed in this paper exhibit several distinct syntactic patterns that are different compared to standard Dutch, posing unique challenges for parsing models. For this analysis, we base ourselves on the most commonly observed characteristics of the SDD group, as described in Section 2.2.

To evaluate whether the trained models can handle these syntactic variations, we begin by filtering sentences that exhibit these patterns from the entire test set (n = 2,616). This allows us to create smaller partitions focused on specific syntactic variations for targeted evaluation. We examine sentences that feature the following patterns: subject doubling (n = 162), negation doubling (n = 43), *en* negation (n = 24), deviating comparative conjunctions (n = 30), expletive *dat* (n = 35) and deviating clause introductions (n = 14).

For clarity and comprehensiveness, this discussion is limited to the overall best-performing models from Section 3. Specifically, we focus on the BERTje-based model across all three dataset partitions and evaluate its performance on sentences exhibiting specific syntactic patterns. Table 2 presents the results of model performance and its improvement across different training setups.

Overall, we observe a consistent and significant improvement across all categories when comparing the standard Dutch parser to the dialect-tuned models. As in the main experiments, a combination of standard and dialect Dutch leads to the highest overall performance. The most notable gains occur in areas affecting negation – specifically, negation doubling and *en* negation. This suggests that while these syntactic irregularities pose challenges for normative language models, it is possible to align them more effectively to handle such data with minimal additional training resources.

An additional challenge for standard models arises in processing alternative comparative conjunctions. As discussed in Section 2.2, certain dialect varieties use markers such as *of* and *als* – typically reserved for disjunctions and conditions respectively – as comparative elements. Unsurprisingly, models trained on normative Dutch struggle with parsing and interpreting sentence structure when these markers are present. However, as with the cases described above, incorporating even a relatively small number of such instances significantly improves the model's ability to handle this syntactic variation more effectively.

4.2 Geographical Variation

The final part of our discussion consists of a formal analysis of the systems per larger dialect group. We split the test set into four smaller partitions. Each partition consists entirely out of sentences from one of the four major dialect varieties in the corpus: West Flemish, East Flemish, Zeeland Flemish and Brabantic. Note that border cases are resolved according to the current provincial/national borders. Note that dividing by provincial borders is not optimal, as many nuanced border cases exist where dialects blend across regions. However, for this preliminary analysis, such a division provides a practical and sufficiently clear framework.

Following the approach outlined in the previous section, we evaluate the best-performing models from each training set (Standard, Dialect and Merged) on the newly created dialect-specific test partitions. This analysis allows us to determine whether certain parsers are better equipped to handle specific dialects or if performance varies across different dialect groups. By comparing results across these partitions, we gain insight into the models' ability to generalize across dialectal variation. Table 3 presents the performance of each model on the various dialect groups.

As can be seen from the table, overall performance per dialect is consistent with the earlier obtained scores in Section 3 for the dataset as a whole. Once again, models trained on a combination of standard Dutch (Lassy + Alpino) and dialect (GCND) data perform best, which is consistent for each of the four evaluated dialect groups. It should be noted, though, that East Flemish performs markedly worse than the other language va-

Pattern	Dataset	UCM	LCM	UAS	LAS
Subject Doubling	Standard	52.47	20.99	81.91	72.49
	Dialect	60.49	40.12	88.36	82.77
	Merged	67.90	46.30	89.85	84.83
Negation Doubling	Standard	34.88	23.26	74.44	65.31
	Dialect	55.81	41.86	85.99	80.93
	Merged	58.14	46.51	86.95	82.73
En Negation	Standard	21.74	8.70	67.89	53.66
	Dialect	52.17	39.13	84.82	78.60
	Merged	65.22	47.83	90.31	85.27
Expletive Dat	Standard	31.43	24.81	80.38	71.63
	Dialect	28.57	25.71	85.55	82.05
	Merged	42.86	34.29	88.01	84.39
Comparative Conjunctions	Standard	31.03	17.24	69.90	62.46
	Dialect	34.48	24.14	79.37	72.70
	Merged	44.83	37.39	83.18	79.51
Clausal Introduction	Standard	42.86	14.29	83.44	69.94
	Dialect	50.00	21.43	86.36	76.70
	Merged	57.14	28.57	87.71	80.45

Table 2: Performance of the BERTje-based model on syntactic pattern-specific sentences across dataset partitions

rieties in all three training setups. We hypothesize two possible explanations for this phenomenon.

First, the training data for the underlying BERTje encoder (De Vries et al., 2019) consists entirely of Dutch as spoken in the Netherlands (i.e., non-Flemish), which may make it more challenging for the model to align with southern Flemish dialects, leading to lower scores. This hypothesis initially seems plausible, given the high performance on Zeeuws (a dialect native to the Netherlands) and Brabants (generally considered the Flemish dialect most similar to Netherlandic Dutch). However, the performance of West Flemish does not fully support this explanation. Further experimentation with the robbert-2023 model, which was pre-trained on both (dialect) Flemish and standard Dutch, does not yield significant improvements on the East Flemish dataset. The performance gap between East Flemish and the other dialects therefore remains consistent, regardless of the encoder used.

A more plausible explanation lies in the specific syntactic patterns characteristic of each dialect group. Analyzing the distribution of these syntactic features, we find that in the East Flemish test set, 14% of the sentences contain one or more of the patterns discussed in Section 2.2. This is the highest proportion among all dialect varieties (West Flemish: 13%, Zeeland Flemish: 4%, Brabantic: 8%) and may account for the model's decreased performance. A closer examination reveals that East Flemish has the highest proportional occurrence of double negation markers – almost double that of West Flemish – as well as a high frequency of *en* negation. These are precisely the syntactic patterns on which the standard models struggled (see Section 4). Although, as shown in Table 3, performance improves somewhat with fine-tuning on dialect data, the average scores for these patterns remain among the lowest in the experiments.

5 Ablation Studies

While the four main dialect groups in this paper have many lexical and syntactic commonalities, they are still ways in which they are notably distinct. Our approach of training on a combined dataset that includes all dialects may unintentionally obscure important dialect-specific characteristics. A potential problem herein is that the model is encouraged to generalize across shared syntactic patterns rather than capturing fine(r)-grained variations. Therefore, an additional set of experiments of training and evaluating performance on individual dialects is crucial to understanding how well the model can preserve these finer linguistic distinctions.

Concretely, we divide our original training, development, and test sets into four smaller subsets, each corresponding to one of the four main dialect groups: West Flemish, East Flemish, Zeeland Flem-

Training Dataset	Dialect group	UCM	LCM	UAS	LAS
Standard	*Brabantic*	62.22	49.10	81.45	75.71
	Zeeland Flemish	65.14	50.00	84.45	77.26
	West Flemish	64.42	49.04	84.64	78.48
	East Flemish	58.90	43.78	78.65	71.64
Dialect	*Brabantic*	73.76	62.44	89.62	85.34
	Zeeland Flemish	70.41	56.88	89.00	82.93
	West Flemish	71.39	58.41	89.14	84.67
	East Flemish	69.54	53.86	86.09	80.64
Merged	*Brabantic*	77.15	68.10	90.50	87.24
	Zeeland Flemish	75.00	62.84	89.58	84.72
	West Flemish	75.00	64.54	90.79	87.20
	East Flemish	71.22	58.90	86.60	82.52

Table 3: Performance on the partitioned test set for each of the dialect varieties using various training datasets

ish, and Brabantic. Table 4 provides an overview of the dataset sizes after partitioning, detailing the distribution across the new training, development, and test sets for each group.

	# Train	# Dev	# Test
West Flemish	7,034	820	833
East Flemish	6,895	886	893
Zeeland Flemish	3,287	418	436
Brabantic	3,574	480	422

Table 4: Number of sentences present in the training, development and test sets for each of the dialects

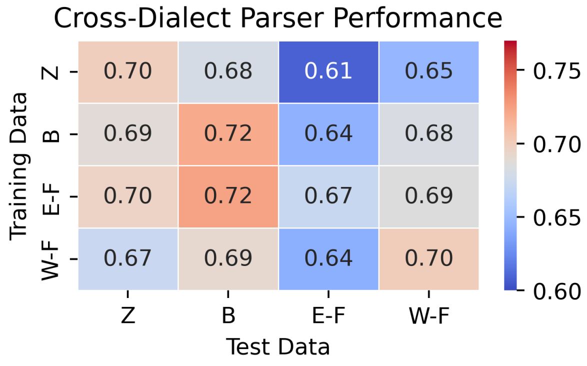

Figure 3: Cross-Dialect results for West Flemish (W-F), East Flemish (E-F), Brabantic (B) and Zeeland Flemish (Z)

Next, we train four dialect-specific parsers, each using the corresponding partitioned dataset. For each dialect, we fine-tune a parser with the optimal encoder identified in Table 1 (BERTje) and evaluate its performance in two ways: (1) on its respective dialect-specific test set and (2) on the other individual test sets to assess generalization across dialects. In order to provide an interpretable overview, Figure 3 displays a comparative heatmap indicating cross-dialect performance. The scores presented in the figure are the average of the four metrics that were used earlier. A full overview, including all metrics can be found in the Appendix.

From the figure, we infer that irrespective of the amount of available training data, models trained on a particular dialect are consistently the best performing model on that specific evaluation data. We also find that performance for the East Flemish dialect is notably worse, which is consistent with our findings from Section 4.2.

Interesting however, is the fact that the East Flemish model itself generally performs on-par or slightly below the top models for the other dialects. We hypothesize two possible explanations for this. First, the amount of available training data is highest for East Flemish as a whole. The good performance of the model might therefore be explained simply by the fact the model had access to more training data, resulting in better alignment. We also note here that the West Flemish model, which contains a comparable amount of training data, tends to be the second best overall model, supporting this hypothesis. Another plausible explanation is that, due to the sheer number and remarkable diversity of divergent syntactic patterns present within the East Flemish dataset (as illustrated in Figure 2), the model may have developed a broader proficiency in handling non-normative linguistic structures in general. Rather than merely adapting to individual irregularities, the model could be refining its capacity to process and generate language that deviates from standard norms, thereby demonstrating an overall advantage when working with non-standard linguistic forms. This suggests that exposure to a wide array of syntactic variations enhances the model's flexibility in parsing, interpreting, and predicting structures that fall outside the boundaries of conventional or prescriptive grammar.

6 Conclusion

This paper has explored the difficulties and potential of dependency parsing for SDDs, a linguistically diverse yet underrepresented area. By using a lexically standardized corpus, we focused on syntactic variation across four key dialect groups: West Flemish, East Flemish, Brabantic, and Zeeland Flemish. Our results show that models trained on standard Dutch perform poorly on dialectal input, especially when dealing with ambiguous function words and region-specific structures. Adding dialect data to training significantly boosts performance, though some constructions remain difficult to parse accurately. Our findings suggest that more dialect-sensitive approaches in syntactic modeling could improve parsing accuracy and support the development of more adaptable NLP tools.

References

Giuseppe Attardi, Daniele Sartiano, and Maria Simi. 2021. Biaffine dependency and semantic graph parsing for enhanceduniversal dependencies. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 184–188.

Julia Bacskai-Atkari. 2020. German v2 and doubly filled comp in west germanic. *The Journal of Comparative Germanic Linguistics*, 23(2):125–160.

Sjef Barbiers. 2009. Locus and limits of syntactic microvariation. *Lingua*, 119(11):1607–1623.

Sjef Barbiers, Leonie Cornips, and Jan Pieter Kunst. 2007. The syntactic atlas of the dutch dialects (sand): a corpus of elicited speech and text as an online dynamic atlas. *Creating and Digitizing Language Corpora: Volume 1: Synchronic Databases*, pages 54–90.

Sjef Barbiers, Johan Van der Auwera, Hans Bennis, Eefje Boef, Gunther De Vogelaer, and Margreet Van der Ham. 2005. *Syntactic atlas of the Dutch dialects*, volume 2. Amsterdam University Press.

Gosse Bouma and Gerardus van Noord. 2017. Increasing return on annotation investment: The automatic construction of a universal dependency treebank for dutch. In *Proceedings of the nodalida 2017 workshop on universal dependencies (udw 2017)*, pages 19–26.

Anouck Braggaar and Rob Van Der Goot. 2021. Challenges in annotating and parsing spoken, code-switched, frisian-dutch data. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 50–58.

Anouck Braggaar and Rob van der Goot. 2021. Creating a universal dependencies treebank of spoken frisian-dutch code-switched data. *arXiv preprint arXiv:2102.11152*.

Anne Breitbarth, Melissa Farasyn, Anne-Sophie Ghyselen, Lien Hellebaut, Frederic Lamsens, Katrien Depuydt, Jesse de Does, Jan Niestadt, and Koen Mertens. 2024. Gesproken corpus van de zuidelijk-nederlandse dialecten. 1st release october 2024.

Anne Breitbarth, Melissa Farasyn, Anne-Sophie Ghyselen, and Jacques Van Keymeulen. 2020. Het gesproken corpus van de zuidelijk-nederlandse dialecten. *Handelingen van de Koninklijke Zuid-Nederlandse maatschappij voor taal-en letterkunde en geschiedenis*, 72.

Anne Breitbarth and Liliane Haegeman. 2014. The distribution and interpretation of preverbal *en* in flemish. In Theresa Biberauer and George Walkden, editors, *Syntax Over Time: Lexical, Morphological, and Information-Structural Interactions*, pages 35–56. Oxford University Press, Oxford.

Anne Breitbarth and Liliane Haegeman. 2015. 'en' en is níet wat we dachten: A flemish discourse particle. In *Proceedings of Moscow Syntax and Semantics (MOSS) 2*, volume 78, pages 41–55, Cambridge, MA. MIT Working Papers in Linguistics.

Joan Bresnan, Ronald M Kaplan, Stanley Peters, and Annie Zaenen. 1987. Cross-serial dependencies in dutch. In *The formal complexity of natural language*, pages 286–319. Springer.

Johan De Caluwe. 2009. Tussentaal wordt omgangstaal in vlaanderen. *Nederlandse taalkunde*, 14(1):8–25.

Johan De Caluwe, Steven Delarue, Anne-Sophie Ghyselen, and Chloé Lybaert, editors. 2013. *Tussentaal: Over de talige ruimte tussen dialect en standaardtaal in Vlaanderen*. Studia Germanica Gandensia. Academia Press, Gent. Spieghel Historiael, themanummer.

Daniël de Kok and Tobias Pütz. 2020. Self-distillation for german and dutch dependency parsing. *Computational Linguistics in the Netherlands Journal*, 10:91–107.

Georges De Schutter. 2015. Meervoudige negatie en paardekooper z'n begrip" stapeling". wat heeft de rnd te bieden? *Verslagen & Mededelingen van de Koninklijke Academie voor Nederlandse Taal en Letteren*, 125(3).

Gunther De Vogelaer and Magda Devos. 2008. On geographical adequacy, or: How many types of subject doubling in dutch. *Microvariation in syntactic doubling*, 36:251–276.

De Vogelaer, Gunther. 2006. *Subjectmarkering in de Nederlandse en Friese dialecten*. Ph.D. thesis, Ghent University.

Wietse De Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. Robbert: a dutch roberta-based language model. *arXiv preprint arXiv:2001.06286*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Timothy Dozat and Christopher D Manning. 2016. Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734*.

Melissa Farasyn, Anne-Sophie Ghyselen, Jacques Van Keymeulen, and Anne Breitbarth. 2022. Challenges in tagging and parsing spoken dialects of dutch. *Journal of Historical Syntax*, 6(4-11):1–36.

Anne-Sophie Ghyselen. 2015. 'stabilisering' van tussentaal? het taalrepertorium in de westhoek als casus. *Taal & Tongval*, 67(1):43–95.

Anne-Sophie Ghyselen, Jacques Van Keymeulen, Melissa Farasyn, Lien Hellebaut, and Anne Breitbarth. 2020. Het transcriptieprotocol van het gesproken corpus van de nederlandse dialecten (gcnd). *BULLETIN DE LA COMMISSION ROYALE DE TOPONYMIE & DIALECTOLOGIE (PRINTED)= HANDELINGEN VAN DE KONINKLIJKE COMMISSIE VOOR TOPONYMIE & DIALECTOLOGIE*, 92:83–115.

Liliane Haegeman and Raffaella Zanuttini. 1996. Negative concord in west flemish. *Parameters and functional heads. Essays in comparative syntax*, (3):117–197.

Lien Hellebaut, Anne-Sophie Ghyselen, Melissa Farasyn, Anne Breitbarth, Veronique De Tier, and Jacques Van Keymeulen. 2021. Stemmen uit het verleden: een schat aan informatie voor heemkundigen en andere erfgoedactoren. *HISTORIES MAGAZINE*, (06/07/2021).

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In *Proceedings of the workshop on noisy user-generated text*, pages 9–18.

Anna Jørgensen, Dirk Hovy, Anders Søgaard, et al. 2016. Learning a pos tagger for aave-like language. In *The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Proceedings of the conference*. Association for Computational Linguistics.

Andy Lücking, Giuseppe Abrami, Leon Hammerla, Marc Rahn, Daniel Baumartz, Steffen Eger, and Alexander Mehler. 2024. Dependencies over times and tools (dott). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4641–4653.

Ryan McDonald, Kevin Lerman, and Fernando Pereira. 2006. Multilingual dependency analysis with a two-stage discriminative parser. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 216–220.

Neuckermans. 2008. *Negatie in de Vlaamse dialecten volgens de gegevens van de Syntactische Atlas van de Nederlandse Dialecten (SAND)*. Ph.D. thesis, Ghent University.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200*.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2004. Memory-based dependency parsing. In *Proceedings of the eighth conference on computational natural language learning (CoNLL-2004) at HLT-NAACL 2004*, pages 49–56.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. Maltparser: A data-driven parser-generator for dependency parsing. In *LREC*, volume 6, pages 2216–2219.

Nelleke HJ Oostdijk. 2006. A reference corpus of written dutch. *Corpus design (D-coi 06-01). Persistent identifier: urn: nbn: nl: ui*, pages 22–2066.

Petrus Cornelis (Piet) Paardekooper. 2015. Meervoudige uitdrukking (stapeling) van negatie, vooral in het nederlands. *Verslagen & Mededelingen van de Koninklijke Academie voor Nederlandse Taal en Letteren*, 125(1-2).

Gertjan Postma. 2006. Van'groter dan'naar'groter als' structurele oorzaken voor het verval van het comparatieve voegwoord'dan'. *Nederlandse Taalkunde*, 11(1):2–22.

J de Rooy. 1965. *Als-of-dat: een semantisch-onomasiologische studie over enkele subordinerende conjuncties in het ABN, de Nederlandse dialecten en het Fries, vergelijkend-synchronisch beschouwd*. Ph.D. thesis, Assen: Van Gorcum.

Johan Taeldeman. 2008. Zich stabiliserende grammaticale kenmerken in vlaamse tussentaal. *Taal & Tongval*, 60(2).

Jeroen Van Craenenbroeck and Marjo Van Koppen. 2002. Subject doubling in dutch dialects. In *Proceedings of Console IX*, pages 54–67. Citeseer.

Leonoor Van der Beek, Gosse Bouma, Rob Malouf, and Gertjan Van Noord. 2002. The alpino dependency treebank. In *Computational linguistics in the Netherlands 2001*, pages 8–22. Brill.

Gertjan Van Noord. 2006. At last parsing is now operational. In *Actes de la 13ème conférence sur le Traitement Automatique des Langues Naturelles. Conférences invitées*, pages 20–42.

Gertjan Van Noord, Gosse Bouma, Frank Van Eynde, Daniel De Kok, Jelmer Van der Linde, Ineke Schuurman, Erik Tjong Kim Sang, and Vincent Vandeghinste. 2013. Large scale syntactic annotation of written dutch: Lassy. *Essential speech and language technology for Dutch: results by the STEVIN programme*, pages 147–164.

Willy Vandeweghe. 1971. Om en rond de (om) te-konstruktie. *Studia Germanica Gandensia*, 13:37–41.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big data*, 3:1–40.

A Appendix

Train Dialect	Test Dialect	UCM	LCM	UAS	LAS
Zeeuws	Zeeuws	66.06	49.77	85.54	78.39
	Brabants	62.90	50.68	82.81	76.26
	Oost-vl	56.10	38.63	78.35	70.24
	West-vl	58.23	41.71	83.18	75.44
Brabants	Zeeuws	66.06	49.31	84.18	75.90
	brabants	68.33	53.39	86.10	80.15
	Oost-vl	61.03	42.55	80.79	73.25
	West-vl	62.74	49.04	84.07	77.19
Oost-vl	Zeeuws	66.51	49.31	85.46	77.26
	Brabants	67.87	55.20	86.25	80.26
	Oost-vl	63.72	46.96	82.72	75.80
	West-vl	62.14	48.08	85.22	78.56
West-vl	Zeeuws	64.22	44.72	84.45	75.98
	Brabants	64.93	50.90	83.75	77.18
	Oost-vl	59.57	52.78	81.03	73.75
	West-vl	64.64	49.16	86.42	80.04

Table 5: Full results of ablation experiments.