# LuxBank: The First Universal Dependency Treebank for Luxembourgish

**Alistair Plum, Caroline Döhmer, Emilia Milano,**
**Anne-Marie Lutgen, Christoph Purschke**
University of Luxembourg
Esch-sur-Alzette, Luxembourg
{alistair.plum,caroline.doehmer,emilia.milano}@uni.lu
{anne-marie.lutgen,christoph.purschke}@uni.lu

## Abstract

The Universal Dependencies (UD) project has significantly expanded linguistic coverage across 161 languages, yet Luxembourgish, a West Germanic language spoken by approximately 400,000 people, has remained absent until now. In this paper, we introduce LuxBank, the first UD Treebank for Luxembourgish, addressing the gap in syntactic annotation and analysis for this 'low-research' language. We establish formal guidelines for Luxembourgish language annotation, providing the foundation for the first large-scale quantitative analysis of its syntax. LuxBank serves not only as a resource for linguists and language learners but also as a tool for developing spell checkers and grammar checkers, organising existing text archives and even training large language models. By incorporating Luxembourgish into the UD framework, we aim to enhance the understanding of syntactic variation within West Germanic languages and offer a model for documenting smaller, semi-standardised languages. This work positions Luxembourgish as a valuable resource in the broader linguistic and NLP communities, contributing to the study of languages with limited research and resources.

## 1 Introduction

The Universal Dependencies (UD) project has facilitated the production of treebanks across many languages, although some languages are still not represented almost 10 years after its original release (Nivre et al., 2016). With 161 languages represented as of the latest release, and a total of 283 treebanks across these languages, the language coverage is undeniably vast.[1] The range of languages includes many of the major world languages, as well as varieties and dialects. However, some languages are still not represented at all, and Luxembourgish was one such case until recently.

A West Germanic language closely related to German, Luxembourgish is spoken by roughly 400,000 people, mainly in Luxembourg (Gilles, 2019). Historically, Luxembourg has had a complex multilingual society where French and German have been predominantly used for official and formal (written) communication. In contrast, Luxembourgish was mostly a spoken language used informally between Luxembourgers until recently. With the rise of digital and social media, however, Luxembourgish has started to develop in the written domain and significant amounts of text data have started to become available, coupled with active language policies promoting Luxembourgish. Research in Natural Language Processing (NLP) for Luxembourgish has been limited until now, often in favour of French, German, and English. This has resulted in a situation where Luxembourgish is considered by some to be a 'low-*research*' language, as opposed to a low-*resource* language.

In addition, large-scale syntactic annotation and analysis has not been undertaken before for Luxembourgish, making Luxembourg one of the few countries whose national language is not represented in the UD treebanks. This remains true despite the fact that four treebanks are available for Standard German (Völker et al., 2019; McDonald et al., 2013; Zeman et al., 2018; Basili et al., 2017), as well as three non-standard treebanks for Swiss German (Aepli, 2018), Low Saxon (Siewert et al., 2021) and Bavarian (Blaschke et al., 2024). None of these represent a Middle-German variety, however, indicating an opportunity to extend the coverage for varieties of (or related to) German.

Aiming to address this gap in research, we present LuxBank, the first UD treebank for Luxembourgish. This project will be the first large-scale quantitative analysis of Luxembourgish syntax, and with this paper, we introduce the first formal guidelines for Luxembourgish language annotation. To this end, we present work related to Luxembourgish

[1] Latest release at the time of writing: 15.05.2024.

*Proceedings of the 22nd Workshop on Treebanks and Linguistic Theories (TLT 2024)*, pages 30 - 39
December 5-6, 2024 ©2024 Association for Computational Linguistics

in Section 2 and describe the creation of LuxBank in Section 3, including highlighting notable syntactic phenomena. We discuss difficulties encountered in the creation process in Section 4 and conclude the paper with Section 5.

## 2 Related Work

Four UD treebanks exist for German, GSD (McDonald et al., 2013), PUD (Zeman et al., 2018), LIT (Basili et al., 2017) and the largest, HDT (Völker et al., 2019), at around 189k sentences. For non-standard varieties of German there are three UD treebanks: the UZH for Swiss German (Aepli, 2018), the LSDC for Low Saxon (Siewert et al., 2021) and as of recently, MaiBaam for Bavarian (Blaschke et al., 2024).

Two sets of guidelines for the UD project have been released since its inception, the first for version 1 (Nivre et al., 2016) and the second for version 2 (Nivre et al., 2020). As the current iteration of the project is version 2, we adhered to these guidelines, although we will discuss some aspects of the version 1 guidelines that could have been useful for our project in Section 4.

### 2.1 Luxembourgish Syntax

Early work on the syntax of Luxembourgish can be found in Schanen (1980) and in a few chapters of grammar books (Schanen and Zimmer, 2012). Certain characteristics of Luxembourgish syntax were later on investigated by dialectologists working on syntactic phenomena in West Germanic (Glaser, 2006) or presented in overview papers on Luxembourgish (Gilles, 2023). A more in-depth analysis of syntactic features was conducted by Döhmer (2020), and there are studies on neighbouring topics, namely pronominal reference for female persons (Martin, 2019) and variation in inflectional morphology (Entringer, 2022), but linguistics research on Luxembourgish syntax and on grammar in general is still in its beginnings. As there is relatively little research literature, we will invest more time into detecting, discussing, and categorising syntactic phenomena parallel to the annotation.

### 2.2 Luxembourgish NLP

Luxembourgish is underrepresented in NLP compared to its linguistic neighbours, French and German. Early research includes resources for NLP tasks (Adda-Decker et al., 2008), analysis of writing patterns (Snoeren et al., 2010), and a corpus for language identification (Lavergne et al., 2014). Recent advancements feature sentiment analysis pipelines (Sirajzade et al., 2020; Gierschek, 2022), an orthographic correction pipeline (Purschke, 2020), a zero-shot topic classification approach (Philippy et al., 2024), and automatic comment moderation (Ranasinghe et al., 2023). LUX-ASR provides Automatic Speech Recognition for Luxembourgish (Gilles et al., 2023a,b), while language models like LuxGPT leverage transfer learning from German (Bernardy, 2022). Additionally, LuxemBERT matches multilingual BERT's performance in Luxembourgish tasks (Lothritz et al., 2022, 2023), and ENRICH4ALL supports a multilingual chatbot in administrative contexts (Anastasiou, 2022). While some tools and models exist for basic language processing, such as a limited spaCy integration[2] and the python tool spellux for lemmatisation[3], there is no published work on these tasks.

## 3 LuxBank

In this section, we set out the methodology for the first round of annotations for LuxBank and reflect on specific linguistic conditions, such as standardisation and structural properties of Luxembourgish. The initial steps include translating the Cairo CICLing sentences, setting up preprocessing, as well as defining the annotation process. For the continuation of this project, we present the next steps in section 3.4, which are focused on adding further sentences from various domains of writing.

The project group working on LuxBank is made up of researchers from a range of different disciplines and specialisations: Two PhD researchers from the research project TRAVOLTA[4] with a background in linguistics, one expert for Luxembourgish grammar and syntax, and one computational linguist specialising in NLP for Luxembourgish. This is of central importance to our approach, as we are trying to incorporate computational processing and linguistic analysis on an equal footing in the development of the project. This is also due to the fact that linguistic experts are often underrepresented in computational linguistics projects. In the following, we describe the data annotation and analysis process.

[2] https://github.com/PeterGilles/Luxembourgish-language-resources/blob/master/spaCyforLuxembourgish.ipynb

[3] https://github.com/questoph/spellux

[4] https://purschke.info/en/travolta

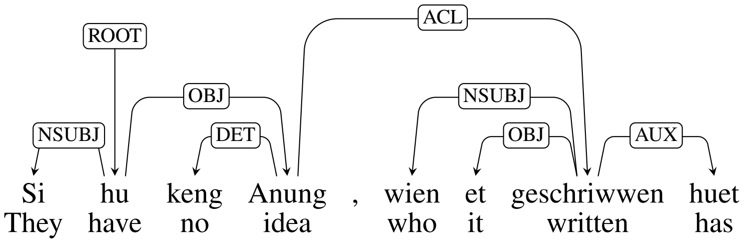

Figure 1: Auxiliary verb in sentence c12.

The Luxembourgish language is not fully standardised and presents a considerate amount of variation, be it lexical, grammatical, or phonological (Entringer et al., 2021). For this project, we decided to use written Luxembourgish according to the official spelling rules.[5] Luxembourgish has an 'emerging standard' and regional variants are being levelled. It is unclear whether there is significant syntactic variation stemming from the different dialects. Given the small size of the country and the ongoing efforts at standardisation, we argue that the variant of written Luxembourgish we are using comes very close to a standard language. The syntactic variation we find in the data is limited and can in most cases be explained through structural reasons.

For our first annotation set, we translate the 20 sentences from the Cairo CICLing corpus into Luxembourgish to ensure comparability. For the second round of annotations we will focus on news texts (journalistic language), as they represent a domain of formal writing and comply with the latest version of the spelling rules published in 2022.[6] The choice of this specific written data is mainly due to practicality reasons, as those texts are easily accessible and offer a good starting point for the project. In the future, we will be open to add texts from different genres to cover a broader range of written language use in practice.

### 3.1 CICLing Sentences

The first 20 sentences are translated from the Cairo CICLing[7] sentences, as recommended in the UD guide for submitting new treebanks.[8] We use the English sentences as source language, and ask native speakers to perform the translations. We employ the available NLP resources for Luxembourgish to perform tokenisation, that is, the available Luxembourgish model for spaCy and spellux for obtaining lemmas.

Of note for our tokenisation is that we split contracted prepositions and determiners manually, which we adopt from Standard German. For the same reason we do not split hyphenated compound words. We deviate from the German guidelines with the determiner *d'*, which does not exist in German, and for which we follow the French standard of tokenising it as *d'*, therefore keeping the punctuation intact.

### 3.2 Annotation

After the corpus selection, the two PhDs working on this project discuss each sentence. The discussion includes analysing the syntactic structure and dependencies by referring to the UD guidelines for German[9] and current work on Luxembourgish syntax (Döhmer, 2020). The analysis starts by annotating the Part-of-Speech (POS) tags for every token. Then, the PhDs adhere to the classic UD process by starting with the main clause, detecting the root and its dependencies with the constituents of the clause. Afterwards, the secondary clause is the main focus of the discussion, looking at the connection with the main clause and its dependencies. Then, as a further step, the two linguists consult the syntactic expert for Luxembourgish to discuss their previous decisions, make additional changes and have a final validation of the dependency annotation.

The difficulties encountered during the annotation process mainly relate to the following reasons: First, the number of people available to work on this project is limited. Since Luxembourgish grammar is not taught in school, finding student assistants who could be trained as annotators is difficult; Second, the two PhDs working on the annotations have limited experience with UD annotation; and

[5] D'Lëtzebuerger Orthografie, Zenter fir d'Lëtzebuerger Sprooch (ZLS) 2022.

[6] D'Lëtzebuerger Orthografie, ZLS 2022.

[7] https://github.com/UniversalDependencies/cairo

[8] https://universaldependencies.org/release_checklist.html

[9] https://universaldependencies.org/de/

|  | hunn<br>(have) | sinn<br>(be) | goen<br>(go) | ginn<br>(give) | kréien<br>(get) | wäert<br>(will) |
|---|---|---|---|---|---|---|
| main verb | + | + | + | + | + | – |
| copula | – | + | – | + | – | – |
| past tense | + | + | – | – | – | – |
| passive voice | – | + | – | + | + | – |
| subjunctive mood | – | – | + | + | – | +/– |
| future tense | – | – | – | – | – | +/– |

Table 1: Functional properties of Luxembourgish auxiliary verbs, adapted from Nübling (2006) by Döhmer (2020).

third, sometimes there is a missing overlap of Luxembourgish grammatical phenomena with the available UD tags.

### 3.3 Special Linguistic Features

In this section, we introduce the syntactic phenomena that need a more thorough explanation, as the tags offered by the UD are not sufficient to cover all the grammatical details unique to the Luxembourgish sentence structure.

#### 3.3.1 The Verbal Domain

We first focus on the verbal domain, describing the categorisation of different functional verb classes during the initial period of the project.

**Auxiliary Verbs** As with most of the Germanic and Romance languages, Luxembourgish has a set of auxiliary verbs to serve different grammatical purposes, such as periphrastic constructions to express the past tense, subjunctive mood, or passive voice. In general, there are six auxiliaries in Luxembourgish, namely *hunn, sinn, goen, ginn, kréien*, and *wäert*, which can also occur as lexical verbs with the meaning of, respectively, 'to have, to be, to go, to give, to get', with the exception of *wäert* ('will') which has a defective paradigm and only works as a function verb. Each of these verbs, when used as an auxiliary, has a specific function, e.g. tense or mood. When used as main verb, these verbs are marked as *root*, while, when used as auxiliaries, they are marked as *aux*, together with modal verbs. Table 1 summarises the functional properties of the Luxembourgish auxiliary system, and Figure 1 shows an annotated sentence from LuxBank.

**Modal Verbs** Like other Germanic languages, Luxembourgish has a set of modal verbs that indicate the modality of the verbal phrase, i.e., if a situation/action is likely, possible, required etc. These are: *kënnen, mussen, sollen, däerfen* and *wëllen*, meaning, respectively, 'can, must, shall, may, want'. Since there is no dedicated tag for modal verbs in the UD, this category too goes under the *aux* tag. In some grammatical descriptions, they are referred to as 'modal auxiliaries' (Barbiers and Van Dooren, 2017). Therefore, in LuxBank grammatical auxiliaries and modal verbs are marked with the same dependency tag. An annotated example from LuxBank is shown in Figure 2.

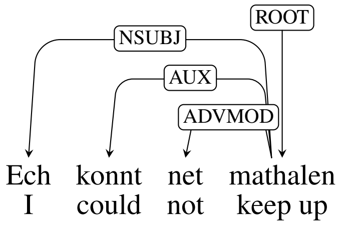

Figure 2: Modal verb in sentence c18.

**Copular Verbs** It is worth underlining here that Luxembourgish, like many other Germanic languages, has more than one verb which can form a copular construction, e.g. *ginn* ('to give') or *sinn* ('to be'). As it is not possible to have more than one copular verb in the UD, at present, *sinn* is registered as copula, while *ginn* is only mentioned as an auxiliary. Figure 3 shows an annotated example from LuxBank.

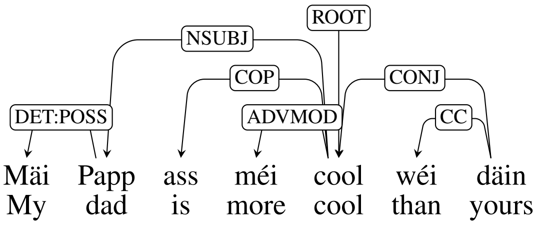

Figure 3: Copular verb in sentence c8.

**Causative Verbs** The verb *doen* 'to do' can be used to form a causative construction. Causatives indicate that a person or event is causing an action

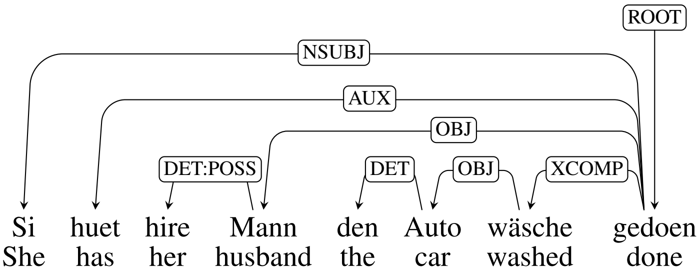
ROOT
NSUBJ
AUX
OBJ
DET:POSS
DET
OBJ
XCOMP
Si huet hire Mann den Auto wäsche gedoen
She has her husband the car washed done

Figure 4: Causative verb in sentence c6.

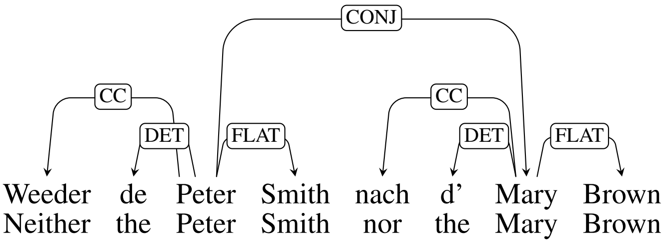
CONJ
CC
DET
FLAT
CC
DET
FLAT
Weeder de Peter Smith nach d' Mary Brown
Neither the Peter Smith nor the Mary Brown

Figure 5: Determiner and proper name in sentence c11.

to happen. This auxiliary was already attested in Old and Middle High German (Hans-Bianchi and Katelhoen, 2011) and persists in Luxembourgish but not in Modern Standard German. However, the use of *doen* is very selective towards its governed verbal phrase, as it can only be combined with specific main verbs. Its status is unclear because it has the functional and structural properties of an auxiliary but the semantic properties of a lexical verb. We tag it as *root* to identify it as a lexical head rather than an auxiliary, considering its limited use and to maintain consistency within the under-specified auxiliary category. An annotated sentence featuring a causative verb is shown in Figure 4.

### 3.3.2 The Nominal Domain

When focusing on further syntactic elements, we find that Luxembourgish also shows a few structural peculiarities in the nominal domain which are worth mentioning.

**Determiner and Proper Name** A common phenomenon in Luxembourgish is the obligatory definite article before proper names. Like in any other noun phrase, the determiner is inflected based on number, gender, and case. Therefore, two or more dependencies in simple noun phrases are quite frequent, especially if the complete name of the person is mentioned. In these cases, we use the tag *det* for the determiner, and following the UD guidelines, *flat* for the second name or surname of the person. The annotated example sentence from LuxBank is shown in Figure 5.

**Possessive Constructions** The genitive is not an active case in the Luxembourgish language. Possessive relations can be expressed with an adnominal dative (only for animate possessors) or with a *vun*-PP (Döhmer, 2020). An annotated example sentence is shown in Figure 6.

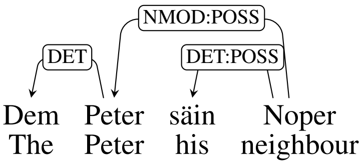

Figure 6: Possessive construction in sentence c7.

### 3.3.3 Other Domains

Since not every phenomenon in Luxembourgish can be analysed with the UD tagset, we decided to use the miscellaneous attributes for the annotation to explicate the phenomena. The miscellaneous attributes, labelled in the MISC column, are intended for the annotators to put in additional information about a tag.[10] At the moment, there are two phenomena that are covered by this tag, the negation and the agreement marker, described in the tag set as *s* clitic.

[10] https://universaldependencies.org/misc.html

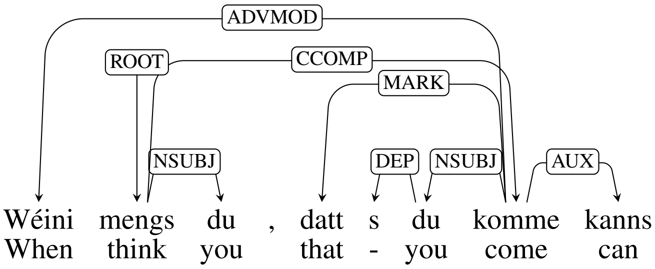
ADVMOD
ROOT
CCOMP
MARK
NSUBJ
DEP
NSUBJ
AUX
Wéini mengs du , datt s du komme kanns
When think you that - you come can

Figure 7: Agreement marker in sentence c14.

**Negation** The negation in Luxembourgish is typically expressed as a negation particle with *net*. In the first version of the UD tagset, the negation was a proper tag, but in the second version the tag is no longer available and is now tagged as *advmod*. We will use the feature *Polarity=NEG* for the negation particle, as is the custom in other UD treebanks.

**Agreement Marker** In subordinate clauses, where the subject is the second person singular (*du/de*), the complementiser is followed by the agreement marker *s*. The *s*-marker is mandatory in this sentence structure and has an orthographically isolated position between the initial element of the subordinate clause and the *du/de*-pronoun (Döhmer, 2020). It developed out of a reanalysis of the inflectional (verbal) *s*-suffix (2nd person singular) and became a clitic before the subject pronoun. Over time it grammaticalised into an obligatory *s*-marker with a fixed syntactic position. As there is no available tag to properly describe this phenomenon, we decided to use the *dep* tag and describe it in the miscellaneous column with *clitic*. In general, this is not a case of clitic doubling as in some West Germanic dialects because the subject pronoun itself is not always used as a clitic. Moreover, the *s*-clitic appears after any element in the complementiser position, not only subordinating conjunctions, but also after interrogative phrases or long prepositional phrases (Döhmer, 2020). Therefore, it should not be linked to the complementiser. Given the fact that it is syntactically bound and very predictable in terms of the sentence type in combination with a specific subject pronoun, attaching it to the verb with the *expl* relation (as per the UD guidelines) would not be justified. Although it doesn't behave like a regular clitic, the *clitic*-tag seems to be the most suitable, because of the strong dependence on the subject pronoun *du/de*. This phenomenon has different structural properties in the Continental West Germanic varieties (it doesn't appear in other standard languages, though) and the terminology may vary in some descriptions (Renkwitz, to appear).

Figure 7 gives an example sentence from LuxBank where this phenomenon is annotated.

### 3.4 Planned Work

Extending the coverage of LuxBank is our primary objective, with the next batch of sentences currently being annotated. This batch comprises 50 randomly sampled sentences[11] from news articles from RTL, the main news broadcaster of Luxembourg. For further extensions, we plan to translate sentences from xSID (van der Goot et al., 2021) to support comparability across further NLP tasks in various languages. While working on this extension, we will also add the morphological features in the initial and future set of sentences.

## 4 Discussion

After applying the UD guidelines and analysing the Luxembourgish sentences, we now discuss practical and theoretical aspects related to the syntactic structure of the 20 CICLing sentences, including under-specified tags and potential challenges when incorporating different languages. Although the CICLing sentences are drawn from simple everyday language, the analysis of such sentences can be quite complex, e.g., when they contain elliptic constructions. Ellipses are a common phenomenon in many European languages, but it is difficult to determine syntactic dependencies, when different parts of the sentence have been elided. Among the 20 CICLing sentences, at least five contain some sort of elliptical structure. As a consequence, CICLing corpus might not be the best starting point for developing new treebanks, since some of the fundamental basic syntactical structures are not as well represented.

[11] Sentences longer than 25 tokens were not considered.

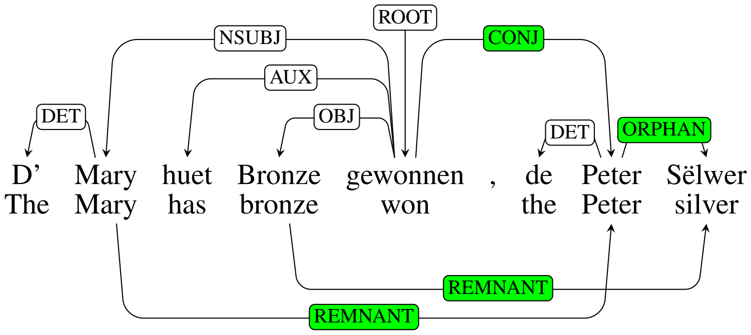
ROOT
NSUBJ
CONJ
AUX
DET
OBJ
DET
ORPHAN
D' Mary huet Bronze gewonnen , de Peter Sëlwer
The Mary has bronze won the Peter silver
REMNANT
REMNANT

Figure 8: UD v2 versus v1 (below) annotation of ellipsis in sentence c9.

To better understand their structure, we analyse the sentences with elliptical structure following both the UD guidelines of version 1 and version 2, see the respective syntactical analysis in Figure 8. Although the version 2 UD guidelines are currently in use, where the dependency between the head of the elliptic sentence and the element depending on the omitted verb is marked as *orphan*, we find the version 1 to be more accurate from a linguistic point of view. In a verb phrase ellipsis, connecting the two *nsubj* under the tag *remnant* and leaving the other dependencies unvaried (i.e. as the verb phrase were there) would better reflect the underlying structure of these sentences.

A further discrepancy between linguistic theories and UD guidelines, as already mentioned in 3.3, concerns the *aux* tag. This tag is under-specified and used for two classes of functional verbs: auxiliaries and modal verbs. While the miscellaneous column can be helpful to deal with the limits of the UD guidelines in practice, it is still a makeshift solution that does not do full justice to phenomena not yet covered by the guidelines. As the feature column is still not enough to distinguish between different verb classes, a dedicated tag to allow better differentiation between auxiliary and modal verbs would be more precise from a linguistic point a view. Moreover, limiting the classification to a single copular verb further reduces the linguistic accuracy of the UD. The possibility to add more than one copular verb would then result in a more realistic representation of the class of copular verbs in Luxembourgish, without compromising the comparability with other languages.

Another aspect regarding the CICLing sentences concerns the modeling of gendered languages. As English usually does not mark the grammatical gender of common nouns, languages with marked gender then need to decide on the grammatical gender of these nouns. Although this is not strictly related to the syntactic dependencies in the sentence, it could lead to a different interpretation and therefore an inaccurate translation of the original sentence. The following example from the CICLing sentences (c7) illustrates this:

(EN) Peter's **neighbour** painted the fence red.

(DE) **Der Nachbar** von Peter hat den Zaun rot (an)gemalt.

(LB) Dem Peter **säin Noper** huet den Zonk rout ugestrach.

As can be seen in the example sentences (marked in bold), even if the grammatical gender is unmarked in English, in both target languages the translators chose the male version of the word, arguably perpetuating the unaware gender bias of male and female roles in society (Bolukbasi et al., 2016). While we do not foresee cases like this in future additions to LuxBank, since we will be using original Luxembourgish material instead of translations, we feel it is important to point this out.

LuxBank is an ongoing project and the main goal is to add more annotated sentences to the treebank. Since this is the beginning of the project, we are continuously adapting the guidelines for Luxembourgish while annotating the data. More linguistic features for Luxembourgish will need to be specified in the future, as they weren't covered in the initial 20 sentences, e.g., loanwords, verb cluster variation, and doubly filled complementisers.

Given the amount of language contact phenomena in Luxembourgish, especially loanwords from German, French, or English are a frequently occurring phenomenon that needs to be addressed. In the nominal domain, further guidelines must be created for French and English compounds, aside from using the *flat* tag, as they are sometimes written as one word, as separate units, or hyphenated, depending on either the spelling norms of the source language or on Luxembourgish orthography.[12] French compounds often appear as multi-word units and are therefore close to syntactic expressions (Goethem and Amiot, 2019). Some of those expressions are directly borrowed into Luxembourgish, e.g. *Projet de loi* 'bill (draft law)' or *Carte d'identité* 'identity card'. These expressions will need to be tagged according to French morphology and left-headedness. It should also be avoided that the French prepositions *de* and *d'* are automatically tagged as Luxembourgish definite articles.

Another common pattern in Luxembourgish syntax is verb cluster variation. The order of elements in 2-, 3-, and 4-verb clusters is variable, when modal verbs or subjunctive auxiliaries appear in subordinate clauses (Döhmer, 2020). In general, word order variation will not affect the deep structure of the sentence, i.e., the dependencies remain the same, but the surface structure will be different. Concerning the left periphery of subordinate clauses, the initial element of the subordinate clause is sometimes extended by a second complementiser, namely *dass/datt* (Döhmer, 2020). Sentences with a doubly filled complementiser, such as *obwuel dass et reent* '(lit.) although that it rains', could cause difficulties in the annotation process because in most cases the complementiser position can only contain a single constituent. All of these phenomena (among others) have to be addressed in the future to develop appropriate guidelines for Luxembourgish.

## 5 Conclusion

In this paper, we introduce LuxBank as the first treebank for Luxembourgish. As the discussion of structural characteristics and challenges encountered when developing annotation guidelines for Luxembourgish show, building a new treebank for a small language represents a theoretical as well as practical challenge. This is particularly true in view of the structural variation in Luxembourgish and its ongoing standardisation. In this context, the decision to bring together a mixed team of linguistic and computational experts has proven crucial to the successful implementation of UD for Luxembourgish.

LuxBank will facilitate a more in-depth understanding of Luxembourgish as a 'low-research' language, making it an invaluable resource not only for linguists but also for language teaching. This treebank project can serve as an aid for spell-checking tools as well as for future grammar checking applications. A tailor-made tagging system derived from earlier versions of LuxBank could ensure higher accuracy and consistency in Luxembourgish text processing and modelling, to help to better organise existing text archives, and to extend the treebank further. In the future, LuxBank will enable easier quantitative exploration of linguistic data, providing insights that were previously more difficult to obtain.

From a typological perspective, it is important to complete the data in the UD treebanks for West Germanic varieties. So far, mainly large standard languages have been incorporated, whereas regional varieties and/or smaller languages are underrepresented. LuxBank adds the first Middle German language description to the UD. This can help to explore syntactic variation and to understand the structural aspects of these languages.

LuxBank will also be beneficial for NLP research and text processing in general. Presently, the support for Luxembourgish is limited to certain tasks (lemmatisation, POS), and the available resources do not use the UD tagset for POS tagging. Building a dedicated treebank for Luxembourgish will make it possible to extend the support for the language in industry-standard tools like *spaCy* to the grammatical level and to offer a comparable tag set for the analysis of syntactic structures. In doing so, LuxBank is laying the foundation for a better representation of Luxembourgish in NLP, both for further research and for the development of customized tools and pipelines.

Luxembourgish can also serve as a model case for describing other small languages and varieties, as these often possess unique characteristics – and resulting challenges – like those discussed in this paper: a limited amount of available resources, a small number of trained linguistic experts, a high amount of linguistic variation (be it lexical, grammatical, or orthographic), a structural influence from other (standard) languages, and a complex multilingual language situation. With this contribution, we aim to position Luxembourgish as a valuable resource for comparable language situations. We also hope to highlight the importance of foundational research for small and non-standardised languages to preserve linguistic diversity in the digital age and make it more visible in NLP.

[12] D'Lëtzebuerger Orthografie, ZLS 2022.

## Limitations

The work presented in this paper is still in progress, and subsequent modifications may be made as the project evolves. It is important to note that finding and recruiting domain experts for data annotation is challenging. Additionally, the amount of variation within the language sometimes makes it difficult to reach a consensus on the classification of phenomena, which has introduced additional complexity to our research.

## Ethics Statement

All data used in this project is freely available and obtained from publicly accessible sources. The human annotators involved in this project were fully compensated for their contributions, as this work forms part of their regular employment responsibilities. Additionally, all data is appropriately licensed for the intended use in this research, ensuring compliance with legal and ethical standards. This adherence to ethical guidelines ensures the integrity and responsible conduct of our research.

## Acknowledgements

This research was supported by the Luxembourg National Research Fund (Project code: C22/SC/117225699).

## References

Martine Adda-Decker, Thomas Pellegrini, Eric Bilinski, and Gilles Adda. 2008. Developments of "Lëtzebuergesch" Resources for Automatic Speech Processing and Linguistic Studies. In *Proceedings of LREC'08*, Marrakech, Morocco. ELRA.

Noëmi Aepli. 2018. Parsing approaches for swiss german. Master's thesis, University of Zurich.

Dimitra Anastasiou. 2022. ENRICH4ALL: A First Luxembourgish BERT Model for a Multilingual Chatbot. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 207–212, Marseille, France. ELRA.

Sjef Barbiers and Annemarie Van Dooren. 2017. Modal Auxiliaries. In Martin Everaert and Henk C. Van Riemsdijk, editors, *The Wiley Blackwell Companion to Syntax*, 2nd ed. edition. Wiley, Hoboken, NJ, USA.

Roberto Basili, Malvina Nissim, and Giorgio Satta. 2017. Toward a treebank collecting german aesthetic writings of the late 18th century. In *Proceedings of CLiC-it*, volume 11, page 12.

Laura Bernardy. 2022. A Luxembourgish GPT-2 Approach Based on Transfer Learning. Master's thesis, University of Trier.

Verena Blaschke, Barbara Kovačić, Siyao Peng, Hinrich Schütze, and Barbara Plank. 2024. MaiBaam: A Multi-Dialectal Bavarian Universal Dependency Treebank.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4349–4357.

Caroline Döhmer. 2020. *Aspekte der luxemburgischen Syntax*. Current Trends in Luxembourg Studies. Melusina Press.

Nathalie Entringer. 2022. *Vun iwwerfëlltene Bussen bis bei déi beschte Witzer. Morphologische Variation im Luxemburgischen – eine variations- und perzeptionslinguistische Studie*. Ph.D. thesis, University of Luxembourg.

Nathalie Entringer, Peter Gilles, Sara Martin, and Christoph Purschke. 2021. Schnëssen. surveying language dynamics in luxembourgish with a mobile research app. *Linguistics Vanguard*, 7:20190031.

Daniela Gierschek. 2022. *Detection of Sentiment in Luxembourgish User Comments*. Ph.D. thesis, University of Luxembourg.

Peter Gilles. 2019. 39. Komplexe Überdachung II: Luxemburg. Die Genese Einer Neuen Nationalsprache. In Joachim Herrgen and Jürgen Erich Schmidt, editors, *Sprache und Raum - Ein internationales Handbuch der Sprachvariation. Volume 4 Deutsch*, pages 1039–1060. De Gruyter Mouton, Berlin, Boston.

Peter Gilles. 2023. Luxembourgish. In Sebastian Kürschner and Antje Dammel, editors, *Oxford Encyclopedia of Germanic Linguistics*. Oxford University Press, Oxford.

Peter Gilles, Léopold Edem Ayité Hillah, and Nina Hosseini Kivanani. 2023a. Asrlux: Automatic speech recognition for the low-resource language luxembourgish. In *Proceedings of the 20th International Congress of Phonetic Sciences*. Guarant International.

Peter Gilles, Nina Hosseini Kivanani, and Léopold Edem Ayité Hillah. 2023b. Lux-asr: Building an asr system for the luxembourgish language. In *Proceedings of SLT)*.

Elvira Glaser. 2006. Zur Syntax des Lëtzebuergeschen: Skizze und Forschungsprogramm. In Claudine Moulin, editor, *Perspektiven einer linguistischen Luxemburgistik. Studien zu Diachronie und Synchronie*, number 25 in Germanistische Bibliothek, pages 227–246. Winter.

Kristel Van Goethem and Dany Amiot. 2019. Compounds and multi-word expressions in french. In Barbara Schlücker, editor, *Complex Lexical Units*, pages 127–152. De Gruyter, Berlin, Boston.

Barbara Hans-Bianchi and Peggy Katelhoen. 2011. Kann man tun und lassen, was man will? Verben zwischen Lexik und Grammatik. *Estudios Filológicos Alemanes*, 201:75–88.

Thomas Lavergne, Gilles Adda, Martine Adda-Decker, and Lori Lamel. 2014. Automatic language identity tagging on word and sentence-level in multilingual text sources: a case-study on Luxembourgish. In *Proceedings of LREC)*, pages 3300–3304, Reykjavik, Iceland. ELRA.

Cedric Lothritz, Saad Ezzini, Christoph Purschke, Tegawendé François D Assise Bissyande, Jacques Klein, Isabella Olariu, Andrey Boytsov, Clement Lefebvre, and Anne Goujon. 2023. Comparing Pre-Training Schemes for Luxembourgish BERT Models. In *Proceedings of KONVENS*.

Cedric Lothritz, Bertrand Lebichot, Kevin Allix, Lisa Veiber, Tegawende Bissyande, Jacques Klein, Andrey Boytsov, Clément Lefebvre, and Anne Goujon. 2022. LuxemBERT: Simple and Practical Data Augmentation in Language Model Pre-Training for Luxembourgish. In *Proceedings of LREC*, pages 5080–5089, Marseille, France. ELRA.

Sara Martin. 2019. Hatt or si? Neuter and feminine gender assignment in reference to female persons in Luxembourgish. *STUF - Language Typology and Universals*, 72(4):573–601.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of ACL*, pages 92–97.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of LREC'16)*, pages 1659–1666, Portorož, Slovenia. ELRA.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Jan Hajič, Christopher D Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal dependencies v2: An evergrowing multilingual treebank collection. *arXiv preprint arXiv:2004.10643*.

Damaris Nübling. 2006. Auf Umwegen zum Passivauxiliar - Die Grammatikalisierungspfade von GEBEN, WERDEN, KOMMEN und BLEIBEN im Luxemburgischen, Deutschen und Schwedischen. In Claudine Moulin and Damaris Nübling, editors, *Perspektiven einer linguistischen Luxemburgistik. Studien zu Diachronie und Synchronie*, number 25 in Germanistische Bibliothek, pages 171–201. Winter, Heidelberg.

Fred Philippy, Shohreh Haddadan, and Siwen Guo. 2024. Forget NLI, use a dictionary: Zero-shot topic classification for low-resource languages with application to Luxembourgish. In *Proceedings of LREC-COLING*. ELRA and ICCL.

Christoph Purschke. 2020. Attitudes Toward Multilingualism in Luxembourg. A Comparative Analysis of Online News Comments and Crowdsourced Questionnaire Data. *Frontiers in AI*, 3:536086.

Tharindu Ranasinghe, Alistair Plum, Christoph Purschke, and Marcos Zampieri. 2023. Publish or hold? Automatic comment moderation in Luxembourgish news articles. In *Proceedings of RANLP*.

Julia Renkwitz. to appear. The agreement of subclause initial elements in Continental West Germanic: Realizations and explanations. In *Syntax aus Saarbrücker Sicht 6*, ZDL Beihefte. Steiner.

François Schanen. 1980. *Recherche sur la syntaxe du luxembourgeois de Schengen: l'enoncé verbal*. Thèse pour le Doctorat d'État, Paris IV, Paris.

François Schanen and Jacqui Zimmer. 2012. *Lëtzebuergesch Grammaire*. Éditions Schortgen.

Janine Siewert, Yves Scherrer, and Jörg Tiedemann. 2021. Towards a balanced annotated low saxon dataset for diachronic investigation of dialectal variation. In *Conference on Natural Language Processing*, pages 242–246. KONVENS 2021 Organizers.

Joshgun Sirajzade, Daniela Gierschek, and Christoph Schommer. 2020. An Annotation Framework for Luxembourgish Sentiment Analysis. In *Proceedings of SLTU-CCURL 2020*, page 172—176, Marseille. Language Resources and Evaluation Conference (LREC 2020).

Natalie D. Snoeren, Martine Adda-Decker, and Gilles Adda. 2010. The study of writing variants in an under-resourced language: Some evidence from mobile n-deletion in Luxembourgish. In *Proceedings of LREC*, Valletta, Malta. ELRA.

Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanović, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi, and Barbara Plank. 2021. From masked language modeling to translation: Non-english auxiliary tasks improve zero-shot spoken language understanding. *Preprint*, arXiv:2105.07316.

Emanuel Borges Völker, Maximilian Wendt, Felix Hennig, and Arne Köhn. 2019. HDT-UD: A very large Universal Dependencies treebank for German. In *Proceedings of udw, syntaxfest 2019*, pages 46–57.

Daniel Zeman, Jan Hajic, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies*, pages 1–21.