Universal Dependencies for the Alemannic Alsatian Dialects

Barbara Hoff, Nathanaël Beiner, Delphine Bernhard
Université de Strasbourg, LiLPa UR 1339
F-67000 Strasbourg, France
{barbara.hoff,n.beiner,dbernhard}@unistra.fr

Abstract

We present the first corpus of Alsatian Alemannic dialects following Universal Dependencies (UD) guidelines, a project which already covers many of the world's languages. Standard languages are represented to a greater extent than non-standard varieties in UD, and our corpus contributes to closing the gap in the lack of resources for Alsatian dialects by presenting the first UD treebank for these dialects, which are spoken in Northeastern France. Our corpus is annotated both with part-of-speech tags and dependency information, as well as French glosses and German lemmas, containing in total 975 sentences and 19,286 tokens, spanning over various text genres. In this article, we present our data, details of the annotation process, as well as some specific syntactic phenomena which differentiate and situate Alsatian with regards to both Standard German and some other German non-standard varieties. The addition of this corpus to the UD project allows for a higher visibility of the Alemannic Alsatian dialects in linguistic research, and provides a valuable resource for research in many fields, including NLP, syntax and comparative Germanic linguistics.

1 Introduction

The project of Universal Dependencies (UD) (Zeman et al., 2024) has the goal of providing cross-linguistic annotation guidelines and treebanks for linguistic research. As of March 2025, there are 296 treebanks in 168 languages available as part of the UD project. While some non-standard German and Germanic varieties are represented (see Section 2), there is currently no UD treebank for Alsatian, an Upper German dialect spoken in the east of France. To address this gap in research, we present the first UD treebank of Alemannic Alsatian dialects, ranging over different source texts and genres. This article will present our work and our corpus, and describe annotation guidelines for some phenomena found in Alsatian. We will first provide a background for the Alemannic Alsatian dialects and available UD resources for non-standard German varieties (Section 2). We will then provide more information about our data and the annotation process (Section 3), and provide examples of some syntactic phenomena in Alsatian and related annotation decisions (Section 3.3 and 3.4).

2 Alsatian and Related Languages in UD

The terms 'Alsatian', in Alsatian 'Elsässisch, Elsassisch', marginally 'Elsässerdytsch', and in French 'alsacien' are hypernyms which refer to all the Alemannic and Franconian dialectal varieties spoken in the Alsace region in Northeastern France. These terms are used by the Alsatian population itself, whether they speak Alsatian or not.

There is no widely used written standard for Alsatian. Various spelling systems have been developed and proposed, to make it possible for all speakers to write in their own variety of Alsatian with shared grapheme to phoneme rules. However, most speakers are not familiar with these spelling systems, and there is thus a lot of variation in how speakers write Alsatian, depending both on the specific variety they speak and on the degree of influence of French and Standard German spelling (Beiner, 2022).

The linguistic terms used to refer to this group of dialectal varieties, e.g. 'Rhine Franconian' or 'Low Alemannic', have been chosen by linguists in reference to the Alemanni and Frankish tribes who settled in Alsace and in the surrounding areas during the 5th century and who were speaking Germanic languages. It is the descendants of those Germanic varieties which are still spoken in Alsace nowadays, as well as in the surrounding regions in Germany and Switzerland.

As shown in Figure 1, the different Upper Ger-

Proceedings of the 23rd International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2025), pages 10–22

man dialects spoken in Alsace are:

- **Rhine Franconian**, characterised by the use of *Pund, Appel* [pʰʊn̥d], [ɒb̥l̥] instead of *Pfund, Apfel* [b̥fʊn̥d], [ɒb̥fl̥] in the southeast of the isophone,
- **South Rhine Franconian**, with the forms *Haus, Ais* [haws, ajs] (New High German diphthongisation) instead of *Hüs, Ys* [hys, is] kept as monophthongs in the south,
- **Northern and Southern Low Alemannic**, differentiated from one another by the ich-laut pronunciation, respectively [ç] or [ʃ] after front vowels and [x] after back vowels in Northern Low Alemannic, and [x] in all positions in Southern Low Alemannic (e.g. *ich, Büch, fycht* [ɪç, b̥yç, fiçd̥ – ɪx, b̥yx, fixd̥]),
- and **High Alemannic**, with *Chind, Chatz* [xɪnd̥, xɒd̥s] instead of *Kind, Katz* [kʰɪnd̥, kʰɒd̥s] in the north.

Figure 1: The dialectal domain in Alsace and Moselle.

They are currently four Standard German treebanks annotated with UD dependencies: HDT (Borges Völker et al., 2019), GSD (McDonald et al., 2013), PUD (Zeman et al., 2017) and LIT (Salomoni, 2017). In addition, there are four treebanks of High and Low German varieties: Bavarian (Blaschke et al., 2024), Swiss German (Aepli and Clematide, 2018), Luxembourgish (Plum et al., 2024), and Low Saxon (Siewert and Rueter, 2024).

We present the first corpus of texts in the Alemannic Alsatian dialects annotated following UD guidelines. It is the second corpus for Alemannic in general, after the Swiss German corpus.

3 Corpus Data

The data for our corpus comes from different sources, spanning different genres of texts, as summarised in Table 1. The training corpus consisted of texts T1–T4.[1] Most texts are in Northern Low Alemannic, except for T3 and 13, as well as some sentences in T12, T14, T15, and T17, which are in Southern Low Alemannic. There are also some sentences in High Alemannic in T14. T5 to T11 are complete texts, while the other texts are excerpts. In total, the corpus contains 975 sentences and 19,286 tokens.

Texts annotated in phase 2 are all translations, usually from French, and sourced from the ParCoLab parallel corpus (Stosic et al., 2024). Some of the texts in the corpus were professionally translated, while others were translated by a member of the project. Both before and during the annotation process, annotators identified and discussed germanisms in the texts, i.e. forms influenced by Standard German which are not part of the traditional Alsatian norm, but that can be used by some speakers nowadays in what we call the modern Alsatian norm. In the texts annotated by a member of the project, such forms were replaced with an Alsatian form judged more accurate with regards to traditional use, and it is the corrected sentence which was annotated. In texts which were professionally translated and transcripts of spoken language, the original version was kept in the `#text_origin` metadata and the corrected sentence was indicated in the `#text` metadata, as was done in the Low Saxon corpus (Siewert and Rueter, 2024, p. 15,977).

[1] The source text of examples used in this article is indicated by the text identifier (T1–11) and the number of the sentence (for example, *s38* for 'sentence 38') according to the numbering of sentences in our corpus. When there is no relevant example available in the corpus, we used constructed examples and do not indicate a text identifier. For Text 11 (T11), we further specify the date of the specific chronicle we want to refer to.

Text	Text and author	Genre	Sentences	Tokens
Phase 1: Training – double annotation				
T1	D'r Hoflieferant, Gustave Stoskopf	Fiction – Theatre	13	88
T2	Recipes, Office pour la Langue et les Cultures d'Alsace et de Moselle (OLCA)	Cooking – Recipe	17	170
T3	Haut-Rhin Magazine	Non-Fiction – Journalistic	15	121
T4	Alemannic Wikipedia	Wiki	17	188
Phase 2: Translated texts – double annotation				
T5	Monday Tales, Alphonse Daudet	Fiction	179	3,804
T6	The Universal Declaration of Human Rights	Legal	83	2,183
T7	The Decameron, Boccaccio	Fiction	19	483
T8	Pierre and the Wolf, Sergueï Prokofiev	Fiction	65	925
T9	The Prodigal Son, Luke (Steiner and Matzen, 2016)	Fiction – Bible	29	628
T10	The North Wind and the Sun, Aesop (Boula de Mareüil et al., 2018)	Fiction	6	126
T11	Chronicles about the regional languages of France, Michel Feltin-Palas	Non-Fiction – Journalistic	177	4,267
Phase 3: Selected sentences – single annotation				
T12	Linguistic and ethnographic atlas of Alsace	Spoken, Ethnotext	60	2,711
T13	Haut-Rhin Magazine	Non-Fiction – Journalistic	12	249
T14	Miscellaneous literary texts, various authors	Fiction	33	370
T15	Miscellaneous texts, OLCA	Fiction, Non-Fiction, Cooking	19	163
T16	Miscellaneous theatre plays, some translated from French, various authors	Fiction – Theatre	171	1,528
T17	Alemannic Wikipedia	Wiki	60	1,282

Table 1: Source texts, genre distribution, and number of surface tokens and sentences per text in the corpus. Texts T1-T4 were only included in the training batch.

3.1 Tokenisation

The corpus was tokenised using an adapted version of the tokenisation script developed for Bavarian (Blaschke et al., 2024).

The tokenisation was manually checked by the annotators in order to, for example, split contracted forms of a preposition and a determiner (*im* into *i* + *m* 'in the') and to correct mistakes when sentences were wrongly split or merged together. Following German UD rules, as well as annotation decisions for Bavarian (Blaschke et al., 2024, p. 10924) and Luxembourgish (Plum et al., 2024, p. 32), contracted forms consisting of a preposition and a determiner were split (see example above), while hyphenated compounds were not (ex: *Grìeni-Lìnse* 'green lentils' (T2, s9)). Epenthetic consonants and associated punctuation were not split, but merged with the previous word, see section 3.4 for details about their annotation.

3.2 Annotation Procedure

The annotators always worked by correcting automatic pre-annotations in order to streamline the annotation process. In phases 1 and 2, the corpus was pre-annotated using three main methods: UDPipe (Straka, 2018), Mistral Large with prompts,[2] and the trainable parsing service on the Arborator-Grew platform (Guibon et al., 2020).[3] See Bernhard et al. (2025) for more information about the pre-annotation process for this corpus.

In phase 3, the annotations from phrase 1 and phase 2 were used to train a new pre-annotation model using the MaChAmp toolkit (van der Goot et al., 2021). In order to increase the amount of available training data, we also used the test splits of the following existing UD 2.15 corpora: German GSD (McDonald et al., 2013), Bavarian MaiBaam (Blaschke et al., 2024), Swiss German UZH (Aepli and Clematide, 2018) and Luxembourgish LuxBank (Plum et al., 2024). Training targeted the following tasks: part-of-speech (POS), dependency relations and German lemma (when available in the training corpus). The model was then used to obtain pre-annotations for sentences from varied source material (see Table 1, Phase 3). We then selected the sentences to be corrected by the annotators using uncertainty sampling based

[2] https://mistral.ai/fr/news/mistral-large-2407

[3] In phase 1, GPT 3.5 and 4 were also used, as well as different training corpora for ArboratoGrew.

on the label probability for POS, head and dependency relation of MaChAmp's predictions. We retained sentences with the largest average uncertainty and with at least three tokens. The objective here was to annotate sentences with phenomena that are more difficult to annotate automatically.

The annotators were two recently graduated master's students with a background in linguistics, both native speakers of French and Northern Low Alemannic and co-authors of this paper. They were hired and compensated according to local standards.

In phase 1, the annotators were trained on a small corpus (the training batch, T1–T4) to familiarise them to Universal Dependencies guidelines. The corpus was divided into the following batches:

- Phase 1 : 1 batch, with double annotations
- Phase 2 : 6 batches, with double annotation
- Phase 3: 2 batches, with a single annotation each

Dividing the corpus into different batches made it possible to quality check more effectively after discussing annotations for a batch, as well as update the annotation guidelines and correct the annotations done so far. During the annotation process in phase 1 and phase 2, the annotators only had access to their own annotations (blind annotation), in order to minimise bias. The corpus was annotated using the ArboratorGrew platform (Guibon et al., 2020) and meetings were regularly scheduled to discuss disagreements in annotation and difficult cases, and to agree on a final version for the batches with double annotation. Annotation decisions were based on UD guidelines for German, as well as annotation decisions in related varieties, like Bavarian (Blaschke et al., 2024) and Swiss German (Aepli and Clematide, 2018). GrewMatch (Guillaume, 2021) was used to access the relevant treebanks. The UD validation script[4] integrated to ArboratorGrew was used to detect mistakes. The inter-annotator agreements were high for phase 2: POS $\kappa \geq 0.90$, dependency $\alpha \geq 0.88$ (Bernhard et al., 2025).

A GitLab repository was used for storing the annotations after each step (blind annotation by each annotator, validated versions). After each upload, a verification pipeline was automatically launched to obtain a report on potential detectable errors in the annotations using Udapi (Popel et al., 2017), in particular the `MarkBugs` module. The pipeline also generated tables showing a side-by-side comparison of the annotations by the two annotators, to facilitate the identification of disagreements. Another GitLab repository was used to write the annotation guide. Each section of the annotation guide is written in a Markdown file and a pipeline based on Pandoc (MacFarlane et al., 2025) automatically generates an HTML version of the guide after each modification.

Annotation times for each batch are indicated in Table 2 for each annotator (A1[5] = annotator 1, A2 = annotator 2). The discussion time for each batch, i.e. meetings during which the two annotators compared their annotations, discussed and resolved disagreements in order to agree on a final version, varied between 5 and 10 hours. Both the annotation and discussion time for the first batches were considerably longer than later batches, since both annotators first had to familiarise themselves with UD guidelines and establish annotation rules for Alsatian. The annotation period for our corpus took place over a period of about 7 months between September 2024 and April 2025.

Batch	Sentences	Tokens	A1	A2
0	62	567	6h	5h40m
1	88	1,947	16h	12h45m
2	93	1,929	8h	7h45m
3	92	1,933	7h	6h50m
4	83	1,672	6h	6h10m
5	104	2,554	8h30m	6h15m
6	98	2,381	6h30m	5h
7	176	3,272	16h45m	/
8	179	3,031	/	17h20m
Total	975	19,286	74h45m	68h30m
Average	108	2,143	9h20m	8h33m

Table 2: Annotation time and information about the nine annotation batches

3.3 POS Tags and Dependencies

The annotation guide developed by the annotators and other members of the project is available online.[6] This section presents some annotation decisions and syntactic constructions specific to Alsatian. The Alsatian dialects differ phonetically to a large extent from standard German ('fragen'

[4] https://github.com/UniversalDependencies/tools

[5] The number used for each annotator corresponds to the number used in (Bernhard et al., 2025)

[6] French version DOI: 10.34847/nkl.0eac4288 ; English version DOI: 10.34847/nkl.5b6cs6wu.

– ‘fröje, fröwe, froja’), and to a lower extent lexically (‘Kartoffel’ – ‘Grumbeer, Aardepfel’), morphologically (e.g. reduced case system, see the annotation guide) and syntactically (see below). See Tables 4 and 3 in the appendix for details of the statistical distribution of POS tags and dependencies in the corpus.

The same POS tags and dependencies as defined in the German UD guidelines[7] were used to annotate our corpus, with the addition of a few subrelations to add more details. For example, we used the subrelations `:lmod` `:tmod` and `:emph` for the dependencies `advmod`, `obl`, and `nmod`, to indicate when the dependency provided information about location, time, or emphasis. This was not done in other Germanic corpora.

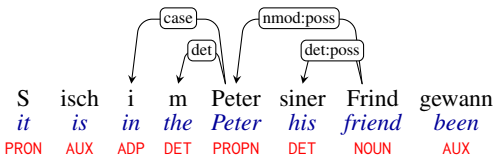
3.3.1 Noun Phrases

Possessive construction Possession is expressed in Alsatian using analytic possessive constructions instead of the genitive case. Two types of constructions are used: either using a prenominal dative, optionally reinforced with the preposition *in*, or using a prepositional phrase starting with *vun* ‘of’ following the possessed.

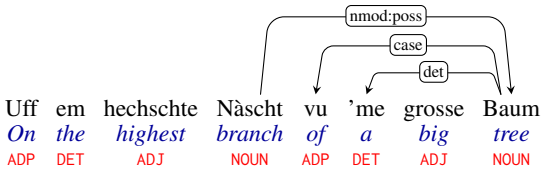
The prenominal dative structure is also found in Bavarian (Blaschke et al., 2024, p. 10926), Luxembourgish (Plum et al., 2024, p. 34) and Low Saxon (Siewert and Rueter, 2024, p. 15979).

We follow annotation guidelines for Bavarian and Luxembourgish and annotate the prenominal dative as follows: in example (1), the possessive pronoun is annotated `det:poss` of the possessed, and the possessor, `nmod:poss` of the possessed. The preposition *in* is annotated as `case` of the possessor, and the determiner, `det` of the possessor. The *vun*-construction (2) is annotated with `nmod:poss`.

(1) Here, *i* ‘in’ is a reinforcement of the dative.
‘It was Peter’s friend.’ (T8, s4)



(2) ‘On the highest branch of a big tree.’ (T8, s3)



Dative case The preposition *in* can also be used as reinforcement of the dative case in other contexts. Below, *Heidelbeere* and *Fräu* are annotated with `obl:arg`, in the same way as dative direct objects, instead of `obl`, as prepositional phrases are usually annotated.

(3) *D’ kleine schwàrze Glickle hàn äu* ***ìn***
The little black eyes have also **in**
nàsse Heidelbeere geglìche.
wet blueberries resembled.
‘The small black eyes resembled also wet blueberries.’ (T5, s170)

(4) *Ich hab ’s* ***in*** *de Fräu gseit.*
I have it **to** the woman said.
Ger. ‘Ich habe es ∅ der Frau gesagt.’

Personal names with a determiner As is also the case in Bavarian (Blaschke et al., 2024, p. 10926) and Luxembourgish (Plum et al., 2024, p. 34), personal names in Alsatian can occur with a determiner. We annotate it as `det` of the name, and use `flat` for personal names with titles (*de Mösiö Hamel* ‘Mister Hamel’ (T5, s3)) and `flat:name` for personal names with first and last name (*de Laurent Lafforgue* (T11/20211123, s6)).

3.3.2 Subordinate Clauses

Relative marker Relative clauses are introduced by the relative marker *wo / wu / wü / wi* in Alsatian. These forms come from Middle High German (MHG) *wâr*[8] ‘where’ that shifted, probably very early, from /ar/ to /u/ then palatalised to /y/ in the Masevaux valley or the Kochersberg, according to Beyer (1964, p. 156). This form is invariant, and Standard German relative markers as *der, die, das* are not used in Alsatian, as is also the case in Bavarian (Blaschke et al., 2024, p. 10926) and Swiss German.

We annotate the relative marker with the POS `PRON` and with the syntactic dependency of the element it replaces. The relative marker is `obj` in example (5).

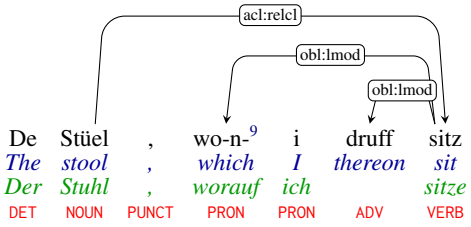
[7] https://universaldependencies.org/de/index.html

[8] *Wo / wu / wü* all come from MHG *wâr*, but the exact etymology of *wi* is unknown, it could be MHG *wie* or *wâr*, see: https://www.woerterbuchnetz.de/ElsWB?lemid=W00044.

(5) *Er hätt gern siner Büch gfillt mit*
he would have like his stomach filled with
***de Schotte, wo** d' Söi gfresse hàn.*
the pods, that the pigs eaten have.
'He would have liked to fill his stomach with the pods that the pigs were eating' (T9, s8)

The invariant relative marker *wo / wu / wü / wi* is also used in cases where a pronominal adverb would be used to introduce a relative clause in German. In such sentences, both the relative marker and the pronominal adverb are annotated with the same dependency as shown in example (6), with the Standard German equivalent in green.

(6) 'The chair on which I sit.'



Pronominal adverbs Similarly to Standard German and other West-Germanic languages, Alsatian has pronominal adverbs. Whereas pronominal adverbs in Standard German can consist of different types of adverbs and prepositions (Pittner, 2024, p. 2–3), they can only contain adverbs starting with *dr-, de-* in Alsatian. They are often reinforced with a preposed *do* 'there' as in *do durich / dodurich*, written merged or split depending on the author. The pronominal character of these adverbs means that they can replace prepositional phrases, although there are some restrictions about the type of preposition phrases they can replace (see Pittner (2024) for more details in Standard German).

Pronominal adverbs can have different functions in Alsatian, and they can modify either a noun (7) or a verb (8–11). They can replace an element which occurs earlier in the sentence (anaphora, 8) or an element which occurs later in the sentence (cataphora, 9). They can also be used to refer to a physical element present in the context of the utterance (deixis, 10).

(7) *S Zìel **devùn** ìsch, d' Rachte ze vernìchte*
The goal **thereof** is, the rights to destroy
With `nmod:poss`. 'Its goal is to destroy the rights.' (T6, s82)

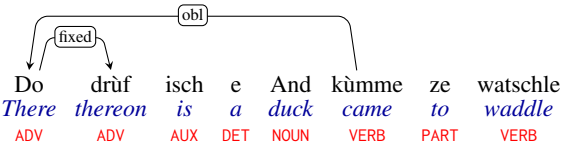
(8) *D' Eh derf nùmme gschlosse ware,*
The marriage may only concluded be,
wànn beidi Hochzitter fréi ùn vollstandi
if both spouses free and fully
***demìt** inverstànde sìnn.*
therewith agreed are.
'The marriage shall be entered into only with the free and full consent of the intending spouse.' (T6, s39)

(9) *ùn wil se sich **defìr***
and because they themselves **therefor**
entschìdde hàn, de soziàle Fùrtschrìtt ze
decided have, the social progress to
férdere ùn besseri Lawersbedìngùnge mìt
promote and better life conditions with
ere greessere Fréiheit ze schàffe.
a bigger freedom to establish.
'And because they have decided to promote social progress and better standard of life in larger freedom.' (T6, s7)

(10) ***Do druf** kànnsch sìtze.*
There thereon can you sit.
Pointing to a chair: 'You can sit on this.'

We annotate pronominal adverbs with the POS `ADV` and the dependency relation `obl` if they modify a verb (8–11), and `nmod` if they modify a noun (7). We chose to use a different dependency than for other adverbs (usually `advmod`) in order to highlight their specificity. We chose `obl` and `nmod` since these dependencies would be used for the prepositional phrase which pronominal adverbs replace, for example compare *s Ziel devùn* and *s Ziel vùn de Tàte*. When the adverb *do* is used to reinforce a pronominal adverb, we annotate the second element as `fixed` of the first one, since it can also be written as one word:

(11) 'Just then a duck came waddling around.' (T8, s6)

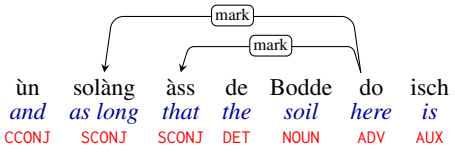


Double/Complex subjunctions In Alsatian, many subordinating conjunctions (subjunctions) that would be formed with one word in Standard German are formed with a preposition followed

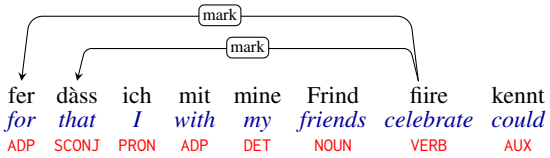
[9] In this example, the relative marker *wo* is followed by an epenthetic *n* to avoid hiatus with the following *i*, resulting in the form *wo-n-i* [vʊni] instead of *wo i* [vʊ.i]. See section 3.4 Epenthesis for more details.

by the subjunction 'dàss/àss'[10] or 'wie', e.g. *fer dàss* (Ger. 'damit'), *for/ebb dàss* (Ger. 'bevor'), *sobàl wie* (Ger. 'sobald') (see Jung, 1983, p. 246). This is also found in Low Saxon (Siewert and Rueter, 2024, p. 15980) and Bavarian (Blaschke et al., 2024, p. 10926). This construction is also found with two subjunctions, for which an additional 'dàss/àss' would not have been needed, and possibly appeared by analogy with the construction preposition + subjunction (Huck et al., 1999, p. 57–60). For example, the following complex/double subjunctions are found: *obwohl àss, trotzdem àss, noochdem dàss, wurum dàss*. We annotate this construction as follows: the first element keeps its original POS, usually ADP (*fer, vor*), ADV (*werum*) or SCONJ (*obwohl, sowyt*), and the second element is always annotated with the POS SCONJ: *dàss / àss / wie*. Both are linked to the subroot with mark.

(12) 'and as long as the soil is here [...]' (T5, s180)



(13) '[...] so I could celebrate with my friends.' (T9, s26)

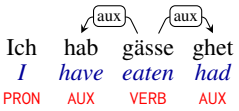


3.3.3 Verb Phrases

Lack of preterite One of the differences between Alsatian and Standard German is the total loss of the preterite tense, which led to new constructions to express it. The habitual aspect is instead expressed using *àls/àss*,[11] which we annotate ADV obl:tmod, as in (14). Because of the loss of the preterite tense, the past anterior is built using the past participles of both the verb and the auxiliary, both annotated using aux (see 15).

(14) *Ich bin* ***ass*** *in de Kinnes mit 'm.*
I am ≈often in the cinema with him.
'I used to go to the cinema with him.' (Jung, 1983, p. 190)

(15) 'I had eaten.'



Lack of future tense The future tense is not grammatically differentiated in Alsatian and the present tense is used to speak about an action that will happen in the future. The verbal form which resembles Future I and II in Standard German (*werden* + infinitive; in Alsatian with *wërre / ware / warde*, see Jenny and Richert (1984, p. 37)) is instead used as a modal verb to indicate an hypothesis or an assumption. In the following example, an assumption is made since the speaker knows Peter's habits:

(16) *Wü isch de Peter? — Är* ***wurd*** *widder im Gaarte hucke.*
Where is the Peter? — He **must** again in the garden hang out.
'Where is Peter? — He must be hanging out in the garden.'

Periphrastic present with *düen* As in some other southern German non-standard varieties, *düen* (Ger. *tun*, Eng. *do*) can be used as an auxiliary in the present tense, although its use is different than in English. In Alsatian, *düen* stresses the active, dynamic nature or the effectiveness of the action, and can also express a prospective mood to carry out the action (Kleiber and Riegel, 2005).

(17) *Sie* ***düen*** *Füesball speele.*
They **do** football play.
Here: 'They are playing football.'

Beneficiary voice Another specificity found in Alsatian is the beneficiary voice (Jenny and Richert (1984, p. 29), see under *middle voice*),[12] used with ditransitive verbs such as *give, steal*. The 'beneficiary' of the verb (in the dative case) becomes the subject, the auxiliary *bekumme/krieje* 'to receive' is conjugated in present tense, and the main verb is in its past participle form (see 18). We annotate these forms with the :bfoc subrelation (beneficiary focus, see UD Voice=Bfoc):

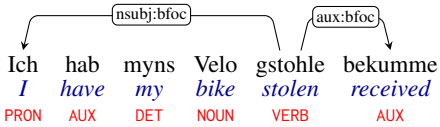
[10] The choice between 'dàss' or 'àss' depends on the variety, the speaker, stylistic choices, as well as the phonetic context.

[11] See meaning 1 in Martin and Lienhart (1899/1907)'s Alsatian dictionary: https://www.woerterbuchnetz.de/ElsWB?lemid=A00487.

[12] The term *middle voice* refers to a different phenomenon in UD guidelines, and we have thus decided to use UD's *beneficiary-focus voice*, which corresponds better to this phenomenon: https://universaldependencies.org/u/feat/Voice.html#Bfoc.

nsubj:bfoc and aux:bfoc, similarly to the annotation of the passive voice.

(18) 'I got my bike stolen.'



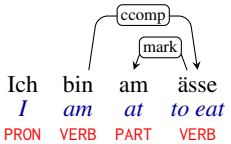
Conditional Mood In Alsatian, the German Konjunktiv II, which we call *conditional*, is not built with *werden* → *würde* as in Standard German, but with the auxiliary *düen* → *dät/dat* (19), or in central dialects with *gan* → *gat*[13] (20). The German Konjunktiv I is only used for the two auxiliaries *hàn* → *héig*, *sin* → *séig* in the varieties of Mulhouse and to the south of the city (Wikiversité, 2023).

(19) *Was* ***dätsch*** *dü mache?*
'What would you do?' (T8, s29)

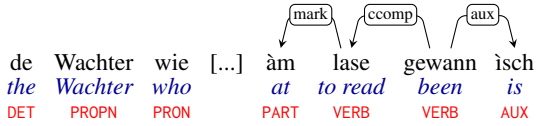
(20) *Wa dü enne g'känn hätsch, d'rno*
If you him known would have, then
wär 's andersch komme, unn d'rno
would be it different came, and then
gäht *'s besser met emm stehn.*
would it better with him stand.
'If you had known him, things would have turned out differently and he would have been better off.'[14]

Progressive Aspect Similarly to other German non-standard varieties, Alsatian expresses the progressive aspect using the present and past tenses. It is built with the auxiliary *sin* 'to be' in the present tense, and a nominalised verb preposed by *am*. We chose to annotate it as a fully grammaticalised construction, thus treating the main semantic verb (e.g. *asse* 'eat') as VERB preposed by *am* PART, the whole group being a ccomp of the auxiliary *sin*, which is annotated as VERB. See (21) for an example in the present tense and (22) in the past tense.

(21) 'I am eating.'



(22) 'The [blacksmith] Wachter, who was reading [the poster with his apprentice].' (T5, s12)



3.3.4 Other Domains

2SG dropped subject pronouns (null subjects) Similarly to Bavarian (Blaschke et al., 2024, p. 10,926) and other non-standard German varieties, second person subject pronouns can be omitted in some contexts in Alsatian (see Hoff (2024a), Hoff (2024b)). There were a few instances of this phenomenon in our corpus (see 23). Since we have not annotated morphological features, the verb in this construction is treated like any other verb, and the absence of a subject pronoun is not annotated for.

(23) ***Brüchsch*** *ken Àngscht ze hàn for mich, papa*
You need no fear to have for me, dad
'You don't need ot be afraid for me, dad.' (T1, s1)

3.4 Use of Features

The only (non-miscellaneous) features used in the corpus were Foreign=Yes and Epenthesis=Yes. See Table 5 in the appendix for details of the statistical distribution of these features in the corpus.

Foreign The feature Foreign was used to indicate loanwords which were not adapted to Alsatian phonology and orthography. For example, the word *Mösjö* 'mister' is adapted, while *Monsieur* is considered a French loanword, not adapted to Alsatian. This feature is always accompanied by the use of the miscellaneous feature Lang, further indicating the language the loanword originates from. In our corpus, some of the foreign languages present were for example: fr (French), de (German), en (English), oci (Occitan), etc.

Proper nouns were treated differently depending on their type: personal names were not annotated as loanwords (ex: *Auguste*), while names of countries (*Bolivie*) or languages (*Creole*) were annotated as loanwords when a non-Alsatian form was used, based on its spelling. Acronyms were annotated as loanwords when they were made up of foreign elements or words. For example, *ONU* '**O**rganisation des **N**ations **U**nies' (in English, *UN – United Nations*) was annotated as a French loanword.

[13] As described by Philipp et al. (1985, p. 5,846), the area extends approximately from Rosheim to the Munster valley. The exact localities can also be found on the map 'täte' in the Wenker (1889/1923) Atlas. See the dark blue backslashes, online at: https://apps.dsa.info/sprachgis/atlas/wa/538.

[14] From the Wenker (1889/1923) Atlas, see question 18 in the locality Bourgheim: https://apps.dsa.info/wenker/transliteration/18607.

Epenthesis The feature `Epenthesis=Yes` was used to indicate words to which an epenthetic consonant was added, for example: *So-**n**-e scheens Hardfir* (T5, s173) 'such a beautiful fire'. An epenthetic consonant, usually *n*, but also *w* in some varieties of Alsatian,[15] is added between two vowels at a word boundary (Wikiversité, 2023). This phenomenon, also called 'Binde-n', is typical of High Alemannic dialects and occurs irrespective of vowel quality (Ortmann, 1998). The type of element/host to which *n* is cliticised plays however a role in its distribution: for example, it never appears on non-finite verb forms (see Ortmann (1998) for more details about this phenomenon and a theoretical explanation). When annotating this phenomenon, we decided to merge the epenthetic consonant to the previous word, as was also done in the Swiss German UD corpus.

Gloss and Lemma The miscellaneous features `Gloss[fr]` and `Lemma[de]` were used to provide word-for-word translations in French and lemmas in Standard German. The gloss in French corresponds to an inflected/conjugated form, while the lemma in German always indicates a 'base' form in the nominative, non-inflected for gender, number, or tense (except for some specific determiners and pronouns, for which gender and number is more relevant). For some words, we indicated multiple German lemma: the first corresponds to the German word with the same etymon, and the second corresponds to modern use in Standard German. For example, *dummel di!* 'hurry!' is annotated `Lemma[de]=tummeln/beeilen`.

4 Conclusion

We have presented our corpus of Alemannic Alsatian dialects annotated following Universal Dependencies guidelines, which is the first one for this dialect. The corpus was pre-annotated using a variety of tools, and annotated by two annotators, while creating and further developing annotation guidelines for Alemannic Alsatian dialects. While many aspects of Alsatian grammar are similar to Standard German, a few specificities were identified and presented in this article. Some syntactic phenomena and annotation decisions for Alsatian were presented and compared to the existing literature and resources on UD corpora for Bavarian, Low Saxon, and Luxembourgish, German varieties related to Alsatian. The corpus described in this paper is undergoing its final review process for addition to the UD repository. It will be available from the following repository: https://github.com/UniversalDependencies/UD_Alemannic-DIVITAL and will add to the resources on non-standard German varieties.

Limitations

Translations Since some of the source texts for our corpus were translated from French, we cannot determine the extent to which the translations were influenced by French and/or Standard German, and to which extent this data differs from naturally occurring Alsatian data. Similarly, annotation decisions were heavily influenced by annotation guidelines for Standard German, which were more accessible and more detailed than guidelines for other non-standard German varieties.

Representation of Alsatian dialects Our project focuses on Low Alemannic dialects and is thus not representative of the whole region: High Alemannic is only represented by a few sentences and Franconian varieties are absent. Furthermore, a great majority of the texts in our corpus are in Northern Low Alemannic. Oral genres and transcriptions are also underrepresented in the corpus, in comparison to written genres.

Ethical considerations

The data used for this project is either freely available from accessible sources, available for research purposes,[16] or the permission to use them for this project was granted by the authors or translators (applies to texts T9 and T11). Excerpts were used in Phases 1 and 3, in accordance with the right to quote. The translators (for texts T5-7-8-11) and the annotators involved in this project were fully compensated for their contributions.

Acknowledgments

This work has been carried out within the framework of the ANR-21-CE27-0004 DIVITAL project supported by the French National Research Agency. We would like to thank the translators organizations and authors who contributed to the

[15] For more details, see Sakumoto (2024).

[16] T10 is licensed under CC-BY-NC-SA and, for the translation of the Universal Declaration of Human Rights (T6), see https://www.ohchr.org/en/human-rights/universal-declaration/universal-declaration-human-rights/about-universal-declaration-human-rights-translation-project

corpus contents: Yves Bisch, Conseil départemental du Haut-Rhin, OLCA, Adrien Fernique, Carole Werner and Michel Feltin-Palas.

References

Noëmi Aepli and Simon Clematide. 2018. Parsing Approaches for Swiss German. In *Proceedings of the 3rd Swiss Text Analytics Conference (SwissText), Winterthur, Switzerland*.

Nathanaël Beiner. 2022. *Quelle(s) norme(s) pour l'écriture de l'alsacien en 2022 ?* Master's thesis, Université de Strasbourg.

Delphine Bernhard, Nathanaël Beiner, and Barbara Hoff. 2025. Pre-annotation Matters: A Comparative Study on POS and Dependency Annotation for an Alsatian Dialect. In *Proceedings of the 19th Linguistic Annotation Workshop*. To appear.

Ernest Beyer. 1964. *La Palatalisation vocalique spontanée de l'alsacien et du badois : sa position dans l'évolution dialectale du germanique continental*. Thèse d'État, Université de Strasbourg.

Verena Blaschke, Barbara Kovačić, Siyao Peng, Hinrich Schütze, and Barbara Plank. 2024. MaiBaam: A multi-dialectal Bavarian Universal Dependency treebank. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10921–10938, Torino, Italia. ELRA and ICCL.

Emanuel Borges Völker, Maximilian Wendt, Felix Hennig, and Arne Köhn. 2019. HDT-UD: A very large Universal Dependencies treebank for German. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 46–57, Paris, France. Association for Computational Linguistics.

Philippe Boula de Mareüil, Frédéric Vernier, and Albert Rilliard. 2018. A Speaking Atlas of the Regional Languages of France. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, page 6, Miyazaki, Japan.

Gaël Guibon, Marine Courtin, Kim Gerdes, and Bruno Guillaume. 2020. When collaborative treebank curation meets graph grammars. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5293–5302, Marseille, France. European Language Resources Association.

Bruno Guillaume. 2021. Graph matching and graph rewriting: GREW tools for corpus exploration, maintenance and conversion. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 168–175, Online. Association for Computational Linguistics.

Barbara Hoff. 2024a. L'omission du sujet référentiel en alsacien, comparée à d'autres variétés germaniques. *Cahiers du plurilinguisme européen*, 16.

Barbara Hoff. 2024b. Reddsch Elsassisch? Referential null subjects in Alsatian, compared to other Germanic varieties. Master's thesis, University of Oslo.

Dominique Huck, Arlette Laugel, and Maurice Laugner. 1999. *L'élève dialectophone en Alsace et ses langues : l'enseignement de l'allemand aux enfants dialectophones à l'école primaire*. Oberlin.

Alphonse Jenny and Doris Richert. 1984. *Précis pratique de grammaire alsacienne: en référence principalement au parler de Strasbourg*. ISTRA.

Edmond Jung. 1983. *Grammaire de l'alsacien, dialecte de Strasbourg avec indications historiques*. Oberlin.

Georges Kleiber and Martin Riegel. 2005. *Les périphrases düen + verbe à l'infinitif en alsacien: Un auxiliaire modal à tout faire*, pages 171–184. John Benjamins Publishing Company.

John MacFarlane, Albert Krewinkel, and Jesse Rosenthal. 2025. Pandoc.

Ernst Martin and Hans Lienhart. 1899/1907. *Wörterbuch der elsässischen Mundarten*. Karl J. Trübner.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

Albert Ortmann. 1998. *Consonant epenthesis: its distribution and phonological specification*, pages 51–76. Max Niemeyer Verlag, Berlin, Boston.

Marthe Philipp, Arlette Bothorel-Witz, and Jean-Jacques Brunner. 1985. Parlers alsaciens. In *Encyclopédie de l'Alsace*, otfried-rhin edition, volume 10, pages 5838–5853. Publitotal.

Karin Pittner. 2024. Pronominal Adverbs in German: A Grammaticalization Account. *Journal of Germanic linguistics*, 36(1):1–31.

Alistair Plum, Caroline Döhmer, Emilia Milano, Anne-Marie Lutgen, and Christoph Purschke. 2024. LuxBank: The first Universal Dependency treebank for Luxembourgish. In *Proceedings of the 22nd Workshop on Treebanks and Linguistic Theories (TLT 2024)*, pages 30–39, Hamburg,Germany. Association for Computational Linguistics.

Martin Popel, Zdeněk Žabokrtský, and Martin Vojtek. 2017. Udapi: Universal API for Universal Dependencies. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 96–101, Gothenburg, Sweden. Association for Computational Linguistics.

Daisuke Sakumoto. 2024. The insertion of voiced labiodental fricative to avoid hiatus in Alsatian: phonological and morphosyntactical conditions of occurence, and diachronical formation process. *Studies on Enunciative Linguistics*, 3:126–153.

Alessio Salomoni. 2017. Toward a Treebank Collecting German Aesthetic Writings of the Late 18th Century. In *Proceedings of the Fourth Italian Conference on Computational Linguistics CLiC-It 2017*, pages 192–197, Torino, Italy. Accademia University Press.

Janine Siewert and Jack Rueter. 2024. The Low Saxon LSDC dataset at Universal Dependencies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15976–15981, Torino, Italia. ELRA and ICCL.

Daniel Steiner and Raymond Matzen. 2016. *D'Biwel uf Elsässisch*. Éditions du Signe, Strasbourg.

Dejan Stosic, Saša Marjanović, Delphine Bernhard, Xavier Bach, Myriam Bras, Laurent Kevers, Stella Retali-Medori, Marianne Vergez-Couret, and Carole Werner. 2024. The ParCoLab parallel corpus and its extension to four regional languages of France. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16014–16023, Torino, Italia. ELRA and ICCL.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Georg Wenker. 1889/1923. *Sprachatlas Des Deutschen Reichs*.

Wikiversité. 2023. Alsacien/grammaire/annexe/synthèse complète — wikiversité. `https://fr.wikiversity.org/wiki/Alsacien/Grammaire/Annexe/Synth%C3%A8se_compl%C3%A8te`. [Online; page available August 22, 2023; accessed February 28, 2025].

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Arofat Akhundjanova, Furkan Akkurt, Gabrielė Aleksandravičiūtė, Ika Alfina, Avner Algom, Khalid Alnajjar, Chiara Alzetta, Erik Andersen, Matthew Andrews, and 633 others. 2024. Universal dependencies 2.15. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, and 43 others. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

A Appendix

Relation	Frequency	Relation	Frequency	Relation	Frequency
punct	3,188 – 16%	nmod	207 – 1%	nmod:lmod	65 – 0%
det	2,207 – 11%	advcl	186 – 0%	det:predet	55 – 0%
case	1,526 – 7%	xcomp	177 – 0%	expl:pv	54 – 0%
nsubj	1,372 – 6%	discourse	176 – 0%	reparandum	44 – 0%
root	975 – 4%	parataxis	168 – 0%	acl	41 – 0%
conj	974 – 4%	appos	153 – 0%	flat:name	31 – 0%
aux	890 – 4%	ccomp	149 – 0%	obl:agent	24 – 0%
advmod	813 – 4%	advmod:emph	141 – 0%	nmod:tmod	23 – 0%
obj	785 – 3%	obl:tmod	138 – 0%	dislocated	23 – 0%
cc	782 – 3%	expl	135 – 0%	csubj	15 – 0%
obl	596 – 3%	aux:pass	128 – 0%	orphan	9 – 0%
amod	572 – 2%	compound:prt	125 – 0%	compound	5 – 0%
mark	542 – 2%	nummod	115 – 0%	cc:preconj	5 – 0%
cop	289 – 1%	obl:arg	111 – 0%	nsubj:outer	2 – 0%
advmod:tmod	286 – 1%	nsubj:pass	108 – 0%	csubj:outer	1 – 0%
obl:lmod	265 – 1%	fixed	104 – 0%	advcl:relcl	1 – 0%
nmod:poss	243 – 1%	flat	103 – 0%	obl:poss	1 – 0%
acl:relcl	212 – 1%	advmod:lmod	81 – 0%	det:preconj	1 – 0%
det:poss	208 – 1%	vocative	66 – 0%		

Table 3: Statistics for dependency relations. For each relation, both the absolute and relative frequency are indicated.

Part-of-Speech Tag	Frequency	5 most frequent tokens
PUNCT – Punctuation	3,188 – 16%	*, . ! : . . .*
NOUN – Noun	2,901 – 14%	*Racht, Mensch, Sproch, Herr, Rachte*
DET – Determiner	2,705 – 13%	*de, m, e, d', s*
VERB – Verb	1,795 – 9%	*het, hàn, gànge, redde, gemàcht*
ADP – Adposition	1,638 – 8%	*vùn, in, ìn, fer, i*
PRON – Pronoun	1,455 – 7%	*wie, mr, mer, se, 's*
ADV – Adverb	1,358 – 6%	*do, so, noch, no, àls*
AUX – Auxiliary	1,348 – 6%	*het, isch, ìsch, hàn, esch*
ADJ – Adjective	1,098 – 5%	*gànz, besser, kleine, fréi, scheen*
CCONJ – Coordinating conjunction	727 – 3%	*ùn, un, odder, oder, Un*
PROPN – Proper noun	413 – 2%	*Peter, Fràankrich, Hamel, Elsàss, Kobüs*
SCONJ – Subordinating conjunction	314 – 1%	*àss, wie, wenn, wànn, dàss*
PART – Particle	311 – 1%	*ze, nit, nìt, net, nitt*
NUM – Numeral	222 – 1%	*zwei, drei, 2, sechs, drissig*
INTJ – Interjection	169 – 0%	*Ja, ja, hein, euh, Jo*
X – Other	43 – 0%	*ta, bon, BA, BE, BI*
SYM – Symbol	11 – 0%	*%, ∗, †, &, n°*

Table 4: Statistics for POS tags. For each tag, both the absolute and relative frequency are indicated, as well as the five most frequent tokens

Feature	Frequency
Foreign=Yes	490 – 2%
Epenthesis=Yes	35 – 0%

Table 5: Statistics over features in the corpus. Both the absolute and relative frequency are indicated.