The IKUVINA Treebank

Mathieu Dehouck

The 𐌀𐌍𐌉𐌖𐌊𐌉 Treebank

Mathieu Dehouck
LATTICE, CNRS, ENS-PSL, Université Sorbonne Nouvelle
mathieu.dehouck@ens.psl.eu

Abstract

In this paper, we introduce the first dependency treebank for the Umbrian language (an extinct Indo-European language from the Italic family, once spoken in modern day Italy). We present the source of the corpus : a set of seven bronze tablets describing religious ceremonies, written using two different scripts, unearthed in Umbria in the XVth century. The corpus itself has already been studied extensively by specialists of old Italic and classical Indo-European languages. So we discuss a number of challenges that we encountered as we annotated the corpus following Universal Dependencies' guidelines from existing textual analyses.

Keywords: Umbrian, Universal Dependencies, Treebank

1. Introduction

The Umbrian language was an Indo-European language from the Italic branch spoken in modern day Umbria (Italy) before the rise of the Roman empire. It is known mostly from seven bronze tablets discovered during the late middle ages known as the Iguvine tablets (or Eugubian, Eugubine tablets). It is one of the best preserved Italic languages after Latin and as such it is of great interest for both the study of old Italic languages and the linguistic environment in Italy at the rise of the Roman empire but also for general Indo-European linguistics. Furthermore, its content sheds light on the religious practices of non Roman, Italic peoples during the last centuries B.C.

The Umbrian language, while being close to Latin, has a number of interesting properties that set it apart, one of them being its wide use of cliticised postpositions where Latin uses prepositions. This could make it useful for research in computational typology for example. There is no fixed orthography in Umbrian and the tablets even use two different scripts which makes it an interesting resource for research in normalisation and/or generalisation techniques to spelling variation. Likewise, the tablets represent various time periods of the language, and thus the various forms, when they are not purely free variations, also represent sound changes that occurred in Umbrian.

Our goal with the IKUVINA treebank is to make the Umbrian language easily accessible for NLP researchers and other interested people. Due to its peculiarities, this corpus can be used for typological, diachronic or normalisation research amongst other.

In this paper, we report on the process of turning an already analysed corpus into CoNLL-U format following Universal Dependencies (Zeman et al., 2022) guidelines. In section 2, we present the Umbrian language, its scripts and the Iguvine tablets. In section 3, we present a number of challenges we encountered as we started to annotate the corpus. In section 4, we discuss the expected output format. Then, we discuss the remaining work and conclude.

Figure 1: The word **ikuvina** as found in the eighth line of the recto of tablet I.

Figure 2: The word "iiovina" as found in the twenty-third line of the recto of tablet VI.

2. The Umbrian Language

The Umbrian language is an Indo-European language of the Italic branch (Hammarström et al., 2021). It was spoken in what is nowadays central Italy around the modern region of Umbria until around the first century B.C. The main Umbrian source is a collection of seven bronze tablets discovered in 1444 near the city of Scheggia (Prosdocimi, 1984). We describe the tablets themselves in section 2.2.

Typologically, Umbrian has a flexible SOV word order supported by a case system akin to the Latin one, with indirect objects often coming after the verb (but not always). It is also a pro-drop language, but the sheer number of imperative verb forms in the corpus (it contains long series of instructions) may not do justice to the actual structure of the language.

2.1. The Scripts

Umbrian was written in both its own Umbrian alphabet (an old Italic script based on the Etruscan alphabet) and in an adapted version of the Latin alphabet at a later stage. Earlier texts written in the original Umbrian alphabet are written from right to left while the ones written using the Latin alphabet are written from left to right. Figure 1 shows the word **ikuvina** written in the original Umbrian alphabet and figure 2 shows its Latin script version "iiovina". Both are forms of the adjective corresponding to the city of Iguvium (modern day Gubbio), from which the English "Iguvine" also derives.

In order to make the distinction clearer, unless stated otherwise, we follow the standard practice of using bold face to render transliterated Umbrian script and standard face with double quotes (when necessary) for Latin script.

One of the peculiarities of the Umbrian alphabet is its lack of dedicated letters for the voiced dental plosive [d] and the voiced velar plosive [g] which are thus rendered by the same characters as their unvoiced counterparts [t] (**t**) and [k] (**k**) respectively. In the later Latin script however, "d" and "g" are used for these voiced sounds, but old practices still occur, thus we find both "crabovie" and "grabovie" (the name of a god) in tablet VI. Note that some earlier [g] rendered as **k** in Umbrian had palatalised by the time of the Latin tablets and where rendered with a plain "i" (Ancillotti and Cerri, 1996), thus giving "iiovina" in figure 2 instead of an hypothetical "*igovina"

While the Umbrian alphabet has a character for the voiced bilabial plosive [b], it is also sometimes written **p** by analogy with the other two plosive series. Note that **p** can also be used to represent a fricative sound which also has its own character in the Umbrian alphabet giving pairs such as **kutef/kutep** (in secret). Thus, the Umbrian **p** can stand for any of the three Latin "b", "p", and "f".

Similarly, the original Umbrian alphabet lacks of an independent character to represent the sound [o] (or [ɔ]), which is usually rendered by the Umbrian character **u** but sometimes by the Umbrian **a**. Ultimately "o" is used in the later Latin script.

However, the Umbrian alphabet has a dedicated letter for [w] (**v**) which merges with [u] (**u**, "v") in Latin versions. And it also has two unique characters, one noting what seems to be a post-alveolar fricative (transliterated **ç**) rendered "š" in later Latin tablets, and one for a kind alveolar fricative trill (transliterated **ř**) rendered "rs" in later Latin tablets.

2.2. The Iguvine Tablets

The seven bronze tablets have sizes ranging from 40 cm × 28 cm for the smallest (tablets III and IV) up to 86 cm × 56.5 cm for the largest (tablets VI and VII) (Weiss, 2019). The seven tablets describe rites and religious ceremonies to be performed by an Umbrian brotherhood including animal sacrifices, purification rituals and food offerings to the gods.

Strong similarities between the Umbrian and the Latin sections of the text and a number of sound changes have led specialists to conclude that the Latin section is a rendering of the same ceremony already described in the Umbrian section but was written at a later stage of the language history (Poultney, 1959).

Table 1 reports on a number of statistics about the tablets broken down by face and scripts. Note that this is only relevant for the verso of tablet V which has inscriptions in both the earlier Umbrian script and the later Latin one.

Tab.	Face	Script	Lines	Chars	Words
I	recto	Umbrian	34	1268	231
I	verso	Umbrian	45	1852	331
II	recto	Umbrian	1+43	1988	323
II	verso	Umbrian	29	1164	198
III	recto	Umbrian	35	1076	177
IV	recto	Umbrian	33	1083	165
V	recto	Umbrian	29	856	154
V	verso	Umbrian	7	146	26
		Latin	11	474	96
VI	recto	Latin	59	4603	844
VI	verso	Latin	65	5800	1020
VII	recto	Latin	54	4443	736
VII	verso	Latin	4	254	43

Table 1: Basic statistics about the raw unannotated Iguvine tablets. The number of lines, characters and words are reported for each tablet broken down by faces and script used for writing. Note that on tablet II, there is a line written vertically in the bottom left corner.

We estimated the number of characters using a standard transcription available in (Poultney, 1959) and on the tablets website[1] ignoring word separators. Since there are a few corrections and what seems to be mistakes and/or omissions, the eventual character count in the annotated corpus will diverge slightly from the raw counts from the tables. Likewise, we report the number of "orthographic words". We rely on word separators and line breaks as much as possible, but we count obvious deviations from these principles as unique words (e.g. on tablet I recto, at line 26, the last letter of the word **pesnim/u** (pray) appears on the following line but we still count the word only once). This gives a bit more than 4300 words overall. However, since Universal Dependencies' format allows us to represent dependency at the syntactic word level (e.g. cliticised adpositions can be handle separately from their host) the eventual token count for the annotated corpus will be higher than the raw word count.

Photographs of the actual tablets, as well as facsimiles, transcriptions, a translation in Italian, an Umbrian vocabulary and a number of other resources can be found on a dedicated website[1].

3. Annotation Process

Due to the singularity of the corpus, we followed a different approach to annotation than for most corpora annotated with dependency trees. The corpus is rather short, yet long enough to teach us something meaningful about its language and long enough to make it worth annotating for NLP practitioners. It has been known for almost six centuries and its language is close from a well documented one (Latin), thus it has already been extensively analysed and many translations have been proposed (all along the same lines). See for example the

[1] `www.tavoleeugubine.it`

work of Bagnolo (1792), Bréal (1875), Poultney (1959), Ancillotti and Cerri (1996). The interested reader can find a much more complete bibliography on the tablets' website[1].

Therefore, the main challenge is not so much to analyse the text itself, but rather to gather the textual analyses that have been published for it and to render them into a machine readable format. In our case we have chosen UD's CoNLL-U format since it is an open format and is widely used and understood by the NLP community. In the following paragraphs, we present a number of challenges that appeared as we annotated the corpus. The proposed solutions are exemplified in table 3.

Note that, while the Umbrian language is fairly well understood, a few words are still obscure and different sources propose different interpretations (see for example (Weiss, 2009) for a discussion on the analysis of the word **erus** for which there is no satisfactory translation yet). For example, **puni** has been understood as mead in (Poultney, 1959) and as flour in (Ancillotti and Cerri, 1996). Therefore, the translations proposed in this paper are tentative and may turn out to be erroneous as we learn more about the ancient Umbrians and their language. The analyses come from (Poultney, 1959) or (Ancillotti and Cerri, 1996), and we rely more on the latter when they disagree since it incorporates more recent works.

3.1. Sentence Segmentation

The original text is segmented into paragraphs. In the sections written in the Umbrian script, vertical spaces and indentations are used, while in the sections using the Latin script, hanging indentation is used. But there is no clear sentence division since punctuation is used for word separation rather than sentence separation.

We thus had to settle on a way to segment the text into sentences. We set the following guiding rule : unless there are some clear indications of subordination, typically a subordinating conjunction (`SCONJ`) such as **pune** (when) or **sve-** (if) sometimes accompanied by an adverb, we try to keep one finite verb per sentence. There are a few exceptions though. On tablet I, for example, we find five almost parallel sentences, they are repeated in table 2, with the verb **fetu** (sacrifice) being repeated twice in the second sentence. We decided to keep it as a unique sentence nonetheless with the second verb coordinated to the first one in order to maintain the original parallelism. The careful reader would have noticed that these sentences seem to come in pairs, the first starting with **preveres** (before the gates) and the second with **pusveres** (after the gates). The missing sentence starting with **preveres treplanes** (before the Trebulian gates), is actually the second sentence of tablet I, but since it is shorter than the other five, and have a different structure, we have not included it in the table.

3.2. Tokenisation

The original text uses punctuation symbols (: in Umbrian, · in Latin) to indicate word boundary. Beside a few cases of missegmentation reported in the literature (e.g. Tablet II, verso, line 20 starts with **pesni:mu:puni:pesnimu** (pray, flour, pray) where the first word should be **pesnimu** without a separator), we followed the original segmentation.

However, many adpositions whose Latin counterparts appear as prepositions, appear as cliticised postpositions (more rarely prepositions) in Umbrian. Since the CoNLL-U format provides a mean to segment surface orthographic words into syntactic words, we have decided to separate cliticised adpositions from their host in the syntactic analysis. We thus analyse **preveres** as **pre veres** (before the gates), **pusveres** as **pust veres** (after the gates), and the common **ukriper** as **ukri per** (for the mount) and **tutaper** as **tuta per** (for the city/state) for example. See table 3 for an example.

We also decided to separate forms made from a subordinating conjunction fused with a pronoun into their original components (e.g. **svepis** as **sve pis** (if someone)).

3.3. Lemmatisation

The main problem regarding lemmatisation is due to the overall small amount of Umbrian data that have reached us. While, thanks to its similarity with other Italic and Indo-European languages, and especially with Latin, it is possible to have a good understanding of the general grammar of Umbrian, we lack many forms for most of the recorded vocabulary. It is therefore virtually impossible to choose a single form (e.g. nominative singular for noun) to be used as lemma for most parts-of-speech. Thus, we have decided to lemmatise closed class words for which we have a better coverage in a first time. After having discussed the question with some of UD's main contributors, we settled on using reconstructed lemma for open class words when necessary and marking such cases with a special `ReconstructedLemma=Yes` feature in the `MISC` column of the CoNLL-U files.

3.4. POS and Morphological Analysis

A few words are ambiguous with regard to their part-of-speech. For example, we find in tablet I the word **vitluf/vitlup** (calf) followed by **turuf/turup** (bull) which would suggest an adjectival use, however we also find a feminine **vitlaf** (heifer) in a very similar context but which is not followed by a noun. The second form could be a case of substantivisation as commonly seen in Latin and in the later Romance languages. Note however, that their Latin cognates "vitulus", "taurus" and "vitula" with the same meanings are usually seen as nouns, so we decided to analyse **vitluf** as an adjective and **vitlaf** as a noun.

Similarly, we find **pustru** (afterward) twice in tablet I ("postro" in tablet VII), which is formally analysed as an adjective in its accusative singular neutral form (Ancillotti and Cerri, 1996), but only appears four times in the whole corpus, each time with an adverbial use, so we decided to mark them as such (`ADV`).

pusveres	treplanes	tref	sif	kumiaf	feitu	trebe	iuvie	ukriper	fisiu	tutaper	ikuvina
preveres	tesenakes	tre	buf		fetu	marte	krapuvi fetu	ukripe	fisiu	tutaper	ikuvina
pusveres	tesenakes	tref	sif	feliuf	fetu	fise	saçi	ukriper	fisiu	tutaper	ikuvina
preveres	vehiies	tref	buf	kaleřuf	fetu	vufiune	krapuvi	ukriper	fisiu	tutaper	ikuvina
pusveres	vehiies	tref	hapinaf		fetu	tefre	iuvie	ukriper	fisiu	tutaper	ikuvina

Table 2: Five parallel sentences occurring on tablet I. They describe sacrifices of animals (pig, cattle and sheep) to be performed around three gates (**preveres**, **pusveres**). The text is not rendered in bold face for readability reasons, but the original is in Umbrian. In the second sentence, the verb **fetu** appears twice, while it only appears once in the other sentences. The first line reads "After the Trebulian gates, sacrifice three pregnant sows to Trebus Jove, for the Fisian mount, for the city of Iguvium.", the other lines are parallels for gods at other gates.

ID	FORM	LEM	UPOS	X	FEATS	H	DREL	DEPS	MISC
1-2	ukriper	_	_	_	_	_	_	_	_
1	ukri	ocar	NOUN	_	Case=Abl|Number=Sing	7	obl	_	_
2	per	per	ADP	_	_	1	case	_	_
3	fisiu	_	ADJ	_	Case=Abl|Number=Sing	1	amod	_	_
4-5	tutaper	_	_	_	_	_	_	_	_
4	tuta	tota	NOUN	_	Case=Abl|Number=Sing	1	conj	_	_
5	per	per	ADP	_	_	4	case	_	_
6	ikuvina	_	ADJ	_	Case=Abl|Number=Sing	4	amod	_	_
7	feitu	fakiom	VERB	_	Mood=Imp|Number=Sing|Person=2 Tense=Fut|VerbForm=Finite	0	root	_	RL=Yes

Table 3: The CoNLL-U format for the sentence **ukriper:fisiu:tutaper:ikuvina:feitu** (Sacrifice for the Fisian mount and the Iguvine city) present on tablet I. X stands for XPOS, H for HEAD, DREL for DEPREL and RL for ReconstructedLemma. Note that we do not use the XPOS column (except for storing annotation during the process) since our corpus is native UD. Note also that not all words already have a lemma (LEM).

Another problem comes from the number of orthographic variants and the tendency for consonants to disappear in word final position. This is well shown in table 2, as there is hardly any doubt that **tre** (three) and **ukripe** (for the mount) in the second line stand for **tref** and **ukriper** respectively. But while the correction is supported by enough evidence in the previous example, for less frequent words, two slightly different forms in different contexts may be fortuitous spelling variants or actual intended different forms. In such case, we stick to the analysis of (Ancillotti and Cerri, 1996) as much as possible.

3.5. Dependency

There are only a few difficulties in the application of the Universal Dependencies' guidelines (Zeman et al., 2022) once we have settled on a morphological analysis. The main reason is likely the small number of complex sentences. The corpus has a number of subordinated clauses but very few relative clauses.

The few subtleties come from ellipsis and direct discourse. We have a case of ellipsis in an enumeration in tablet I as : **tuta:tařinate:trifu:tařinate:turskum:naharkum:numem:iapuzkum:numem** (the Tadinate city, the Tadinate tribe, the Etruscan (name), the Naharcan name (and) the Iapuscan name) where **numem** (name) is elided after **turskum** (Etruscan) and where we therefore attach it directly to the head of the enumeration to maintain symmetry as proposed in UD guidelines. Note that we find the Latin script counterpart of this enumeration in tablet VII and that the Latin version is elided even more as "tuscom·naharcom·iapusco·nome".

Tablet VI contains invocations dedicated to "dei·grabovie" (a tutelary god of Iguvium) with direct report of what ought to be said during the rituals. For example, there are a lot of second person pronouns directed to the god and not to the reader. But there is no specific punctuation distinguishing the direct discourse (directed to the god) from the plain narrative (directed to the reader) thus attachment can sometimes be ambiguous.

4. Output Format

As discussed in section 2, the original corpus has been written with two different scripts (Umbrian and Latin). There has long been a standard transliteration of the Umbrian script using "**ç**" to represent an assumed post-alveolar fricative (rendered "š" in later Latin versions) and "**ř**" for a unique character rendered "rs" in later Latin versions. Therefore we plan on releasing a version of the treebank using the standard transliteration.

However, there exists also an Old Italic block in the Unicode, that is used to encode the Umbrian alphabet amongst other old scripts. Thus we also plan on releasing a version of the section written in Umbrian using the Old Italic block of Unicode to render the original Umbrian script.

Repetition is also an issue. There are a number of very common sentences, for example, **puni:fetu** (sacrifice with flour) and its orthographic variants appear 10 times on tablet I alone. However, we decided to keep each sentence in order to preserve the structure of the corpus and since it is already limited in size. Thus, we will need to address the repetition issue when producing a standard split for training/testing machine learning algorithms.

5. Ongoing and Future Work

Out of the seven tablets, we have annotated most of tablet I and the Umbrian part of tablet V and we are in the process of annotating tablets II, III and IV. The text of tablet I partially annotated, was released in May 2022 as part of the UD 2.10 release (Zeman et al., 2022). Tablets V, VI and VII will appear in following releases.
We also need to find a way to create an interesting standard split (a division in training, development and testing sentences). As we mentioned earlier, there are a few very common sentences and some almost parallel sentences in the Umbrian and Latin sections. This could easily make sentences occur in the various splits and thus make testing metrics artificially high.
As any corpus, the IKUVINA corpus will be subject to evolution if errors are detected or if new discoveries require the corpus analysis to be reevaluated.
When the corpus is completely annotated, a natural research direction will be to see how well models trained on Latin data transfer to Umbrian, and how much work is need to make Latin Umbrian enough to be usable.
Beside Latin and Umbrian, Oscan is another Italic language with a decent amount of materials which could be interesting to the NLP community.

6. Conclusion

In this paper, we have presented the first dependency treebank for Umbrian, an old Indo-European language of the Italic branch. We have presented the source of the corpus : the Iguvine tablets and the scripts they are written with. Eventually, we have discussed a number of challenges appearing when annotating an already analysed corpus from an under-resourced extinct language as well as some solutions we have proposed.

7. Bibliographical References

Ancillotti, A. and Cerri, R. (1996). *Le tavole di Gubbio e la civiltà degli Umbri*. Jama, Perugia.

Bagnolo, G. F. G. (1792). *Le tavole di Gubbio interpretate e commentate*. Torino.

Bréal, M. (1875). *Les Tables eugubines*. Paris.

Hammarström, H., Forkel, R., Haspelmath, M., and Bank, S. (2021). Glottolog 4.5.

Poultney, J. W. (1959). *The Bronze Tables of Iguvium*. American Philological Association, Baltimore.

Prosdocimi, A. (1984). *Le tavole iguvine vol. 1*. L.S. Olschki Firenze.

Weiss, M. (2009). Umbrian erus. *East and West, Papers in IndoEuropean Studies*, page 241–264.

Weiss, M. (2019). *tabulae Iguvinae*. Oxford University Press.

Zeman, D., Nivre, J., and al. (2022). Universal dependencies 2.10. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.