Journal of Open Humanities Data

Introducing STAF: The Saarbrücken Treebank of Albanian Fiction

DATA PAPER

LUIGI TALAMO

ubiquity press

ABSTRACT

The present paper describes the building of STAF, a Universal Dependencies treebank for Albanian. STAF was bootstrapped using a Stanza model trained on previously unreleased data and then manually corrected by three Albanian speakers supervised by the author, who also revised all sentences. STAF focuses on the fiction genre, featuring 200 sentences selected from nine literary texts written by Albanian contemporary authors.

CORRESPONDING AUTHOR:
Luigi Talamo
Language Science and Technology, Saarland University, Saarbrücken, Germany
luigi.talamo@uni-saarland.de

KEYWORDS:
Albanian; treebank; Universal Dependencies; fiction

TO CITE THIS ARTICLE:
Talamo, L. (2025). Introducing STAF: The Saarbrücken Treebank of Albanian Fiction. *Journal of Open Humanities Data*, 11: 3, pp. 1–6. DOI: https://doi.org/10.5334/johd.285

1 OVERVIEW

REPOSITORY LOCATION

https://doi.org/10.5281/zenodo.14552809

CONTEXT

The Saarbrücken Treebank of Albanian Fiction (STAF) is part of a larger effort to build a parallel corpus for typological investigations, the Corpus of Indo-European Prose and more (CIEP+: Talamo & Verkerk, 2022; Verkerk, A. and Talamo, L., 2024), which aims to include the translations of 18 literary texts in 50 languages from 12 families. The parallel corpus is automatically annotated using Stanford Stanza (Qi et al., 2020) with models trained on Universal Dependencies (UD), performing the following Natural Language Processing (NLP) tasks: sentence splitting, tokenization, lemmatization, universal parts-of-speech (UPOS) and universal morphological (Universal Features) tagging, dependency parsing. At the time of writing, twelve languages sampled in CIEP+ do not have pre-trained models and/or available treebanks in the UD collection.

Before the release of STAF in UD v.2.15 (November 2024), Albanian had only one treebank in the UD collection, UD-Albanian_TSA: Toska et al. 2020, which is too small (922 tokens, 60 sentences) to train a reliable model. Since the release of UD v.2.11, a treebank for Gheg Albanian, UD-Gheg_GPS, has been available. Despite the close similarities between this dialect and standard Albanian, which is based on the Tosk dialect but with a complex interaction with Gheg (Camaj, 1984, xv–xvi), the usage of the GPS for annotating written standard Albanian is problematic for a number of reasons. First and foremost, the GPS treebank is based on oral data collected in Kosovo and in Switzerland, containing features of "(semi-)spontaneous speech, like disfluencies and corrections".[1] Furthermore, as is common in oral corpora, the ortography of the GPS treebank is actually the original transliteration of the collected data, reflecting some features of the oral speech and thus differing from the ortography of standard Albanian e.g., *rru:gën*, standard Albanian *rrugën* 'the road.ACC'. Finally, due to the multilingual environment in which the data was collected, GPS contains several examples of code-switching with German, specifically Swiss German.

As for tools performing specific NLP tasks, Kastrati & Biba, 2022 report that almost all of the parts-of-speech and morphological taggers for Albanian are "not available online for NLP purposes". A notable exception is an Albanian model for the Turku Neural Parser Pipeline (Kanerva et al., 2018), which is trained on a large treebank (185K tokens) annotated in the UD framework (Kote et al., 2019); unfortunately, the treebank misses the annotation for dependency parsing and, consequently, the model does not perform this task, which is crucial for both applied and theoretical research.

Finally, a new treebank for standard Albanian has recently been presented in Kote et al., 2024, the Standard Albanian Language Treebank (SALT). SALT is a large treebank (24,537 tokens, 1.4K sentences) annotated in the UD framework with some divergences from the official guidelines (see below) and is used as the 'seed treebank' for bootstrapping STAF. SALT is currently unreleased and its use as the seed treebank for STAF is gratefully acknowledged.

2 METHOD

STEPS

Since CIEP+ is a parallel corpus of literary texts, I have decided to focus on the fiction genre for STAF. Developing resources featuring the fiction genre is particularly useful for cross-linguistic comparison because the vast majority of parallel resources i.e., parallel corpora, only cover the legal and religious genres. Due to its narrative and descriptive nature, the language of the fiction genre is particularly rich both in lexicon and in morpho-syntactic structure; moreover, the fiction genre is often also characterized by a certain amount of dialogue, which to some extent mimics the spoken language.

1 https://gitlab.uzh.ch/uzh-slavic-corpora/gheg-albanian-pear-stories/-/blob/master/README.md.

The following steps were undertaken in the building of STAF: (i) data collection, (ii) automatic processing and (iii) manual correction.

As for the first step, I have legally acquired Albanian books available in digital format. This included full books in various formats (PDF, epub), which were converted to the TXT format using Calibre,[2] as well as free book excerpts offered by on-line vendors. As shown in Table 1, I have sampled 200 sentences from nine fictional books written in standard Albanian by contemporary authors in the 1963–2016 period. The sentence sampling was mostly randomic, but I tried to keep a balance between dialogue and narrative parts. For instance, sentences from Dibra's *Gjumi mbi borë* are quite short and mostly contain dialogue; by contrast, sentences from Qosia's *Një dashuri dhe shtatë faje* and Kongoli's *Ëndrra e Damokleut* contain long narrative and descriptive passages. Finally, some sentences, such as the four sentences from Açka's *Kryqi i harresës*, were chosen in order to cover all the possible merged particles e.g., *t'i* = *të*.PART + *i*.DAT.3SG (see the Reuse potential section).

Table 1 Overview of the texts sampled in STAF.

AUTHOR – TITLE	YEAR	SENTENCES	TOKENS
Ismali Kadare – Gjenerali ï Ushtrisë së Vdekur	1963	42	593
Dritëro Agolli – Njerëz të krisur	1995	11	144
Fatos Kongoli – Lëkura e qenit	2003	12	317
Rexhep Qosja – Një dashuri dhe shtatë faje	2003	36	529
Flutura Açka – Kryqi i harresës	2004	4	64
Fatos Kongoli – Ëndrra e Damokleut	2004	50	1207
Enkelejd Lamaj – Libri i bardhë	2011	16	90
Enkelejd Lamaj – Vendi diku midis	2014	10	229
Ridvan Dibra – Gjumi mbi borë	2016	19	152
Total		200	3325

The second step involved the training of an Albanian model for Stanza using an early (November 2023) version of the SALT treebank (Kote et al., 2024), combined with pre-trained word vectors from the FastText collection.[3] The resulting model was used to bootstrap the annotation of STAF, automatically processing the 200 sentences for tokenization, lemmatization, parts-of-speech and morphological tagging as well as dependency parsing.

In the third step, three Albanian native speakers manually corrected the annotated sentences; two of them, AÇ and RR, are professors for the Albanian language at the University of Tirana (Albania), and the third one, EL, was a student assistant at Saarland University (Germany), who previously received a three-month training in Linguistics and in the annotation of UD treebanks. During a visiting period at Saarland University, AÇ and RR corrected 50 sentences and supervised EL in the annotation of 25 sentences; the remaining sentences were corrected by EL under my supervision. The correction of the annotation focused on the parts-of-speech and morphological tagset and on the dependency parsing, and aimed to correct processing errors as well as annotations that diverge from the UD guidelines. As discussed in Kote et al., 2024, these diverging annotations exist in both the parts-of-speech/morphological tagset and dependency parsing, making SALT non-interoperable with the existing UD treebank (TSA), and hindering its comparability with other treebanks in the UD collection.

QUALITY CONTROL

After manual correction, I performed a full review of all sentences with regard to parts-of-speech and morphological tagsets, as well as dependency parsing. Furthermore, in order to pass the UD validation test,[4] I adapted the annotation of STAF to that of the existing UD treebank, TSA, with some exceptions and new annotation variables. As shown in Table 2, which summarizes

2 Calibre is an open-source software for managing books in digital format, available at https://calibre-ebook.com/.

3 https://fasttext.cc/docs/en/crawl-vectors.html.

4 https://universaldependencies.org/release_checklist.html.

	TSA	STAF	SALT
Multi-word tokens	no	yes	yes
UPOS tags	14	15	17
UPOS for *kam* 'to have' and *jam* 'to be' as copula	AUX	AUX	VERB
Deprels	33	37	32
Dephead for adjectival/nominal predications	adj/noun	adj/noun	copula
Deprel for *për/të* + verb	mark	mark	fixed
Deprel for oblique temporal modifiers	obl	obl:tmod	advmod
Deprel for possessive pronouns	det	det:poss	amod:poss
Deprel for articles of prearticulated adjectives	det:adj	det:adj	det
Deprel for pronominal clitics in clitic doubling	expl	obj/iobj	obj/iobj
Features	36	41	?
Features for adjectives	Gender, Number	Case, Degree, Gender, Number	Case, Degree, Gender, Number
Features for adpositions	–	–	Case
Features for adverbs	Degree	Degree, AdvType	AdvType
Features for articles	Gender	Case, Definite, Gender, Number, PronType	Case, Gender, Number, PronType
Features for possessive markers (*i/e/së/të* + possessor)	Gender	Gender, Number	Case, Gender, Number, PronType
Features for personal pronouns	Gender, Number, PronType	Case, Gender, Number, PronType	Case, Gender, Number, PronType

Table 2 Differences in the annotation of TSA, STAF, and SALT treebanks. UPOS = Universal Parts of Speech; deprel = Dependency Relation; dephead = Dependency Head; features = Universal Features.

the main differences in the annotations of the three treebanks, I have introduced several morphological features, analyzed words as multi-word tokens (MWTs), annotated pronominal clitics as object/indirect objects and added subtypes for dependency relations.[5]

3 DATASET DESCRIPTION

REPOSITORY NAME

Zenodo

OBJECT NAME

STAF

FORMAT NAMES AND VERSIONS

TXT (CoNLL-U format)

CREATION DATES

2023-07-10 to 2024-11-15

DATASET CREATORS

Luigi Talamo (Saarland University), Edita Luftiu (Saarland University), Nelda Kote (Polytechnic of Tirana), Rozana Rushitu (University of Tirana), Anila Çepani (University of Tirana).

5 For a detailed list of the differences between TSA and STAF, please refer to the official UD documentation for Albanian: https://github.com/UniversalDependencies/UD_Albanian-STAF/blob/master/stats.xml.

LANGUAGE

Albanian

LICENSE

Creative Commons Attribution 4.0 International

PUBLICATION DATE

2024-11-15

4 REUSE POTENTIAL

The following are a number of examples of the reuse potential of STAF.

QUANTITATIVE EMPIRICAL RESEARCH

As a validated Universal Dependencies treebank,[6] STAF allows for typological and contrastive studies with respect to over 200 languages. For instance, the annotation of merged particles, which result from the combination of personal pronominal clitics with other personal pronominal clitics and/or subordinator markers, as multi-word tokens i.e., *ma* = *më*.DAT.1SG + *e*.ACC.3SG (see also Toska et al., 2020, 182–183; Kote et al., 2024, 87) has been recently exploited in a cross-linguistic study on pronouns (Talamo et al., submitted).

TRAINING AND DEVELOPING SET

STAF can serve as the dataset for the automatic training of language models used in several Natural Language Processing tasks, such as lemmatization, parts-of-speech tagging and dependency parsing. For instance, Stanford Stanza includes a model for Albanian trained on STAF.

TESTING SET

STAF is manually annotated and carefully checked in each of its annotation field, allowing for its use as a gold standard resource.

ACKNOWLEDGEMENTS

Miss Edita Luftiu worked as the student assistant on this project, annotating and correcting a substantial part of STAF. Mr. Andrew Dryer contributed to the training of Miss Edita Luftiu in Linguistics and in the annotation of UD treebanks. Professor Anila Çepani and Professor Rozana Rushitu helped supervising Miss Luftiu, contributed to the annotation, discussed several annotation choices and explained to me the tricky structures of the Albanian language. They also contributed with teaching materials and on-line resources for Albanian, such as *Kulla e shqipes* (https://gjuhashqipe.com/kulla), a lexicographic and morphological database. Dr. Nelda Kote and her group kindly provided me with their unreleased SALT treebank, which served as the seed treebank for bootstrapping STAF.

FUNDING INFORMATION

This work has been supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through SFB 1102 (Project-ID 232722074) and by Saarland University through two grants: 'UdS-Internationalisierungsfonds: Improving NLP tools for a low-resource language: the Saarbrücken Treebank of Albanian Fiction (STAF)' and 'Towards meaningful coverage of Albanian and Bengali in the Universal Dependencies project'.

COMPETING INTERESTS

The author has no competing interests to declare.

6 Available as UD_Albanian-STAF: https://github.com/UniversalDependencies/UD_Albanian-STAF.

AUTHOR AFFILIATIONS

Luigi Talamo orcid.org/0009-0009-4640-3052
Language Science and Technology, Saarland University, Saarbrücken, Germany

REFERENCES

Camaj, M. (1984). *Albanian grammar: with Exercises, Chrestomathy, and Glossaries* (L. Fox, Trans.). Wiesbaden: Harrassowitz. (Text in English and Albanian, translated from a work originally written in German and Albanian).

Kanerva, J., Ginter, F., Miekka, N., Leino, A., & **Salakoski, T.** (2018). Turku Neural Parser Pipeline: An End-to-End System for the CoNLL 2018 Shared Task. In: *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies.* Association for Computational Linguistics.

Kastrati, M., & **Biba, M.** (2022). Natural language processing for Albanian: a state-of-the-art survey. *International Journal of Electrical and Computer Engineering (IJECE)*, *12*(6), 6432–6439. https://doi.org/10.11591/ijece.v12i6.pp6432-6439

Kote, N., Biba, M., Kanerva, J., Rönnqvist, S., & **Ginter, F.** (2019). Morphological Tagging and Lemmatization of Albanian: A Manually Annotated Corpus and Neural Models. *CoRR*, *abs/1912.00991*. Retrieved from http://arxiv.org/abs/1912.00991

Kote, N., Rushiti, R., Çepani, A., Haveriku, A., Trandafili, E., Meçe, E. K., … Deda, A. (2024). Universal Dependencies treebank for standard Albanian: A new approach. In *Proceedings of the Sixth International Conference on Computational Linguistics in Bulgaria (CLIB 2024)* (pp. 80–89). Sofia, Bulgaria: Department of Computational Linguistics, Institute for Bulgarian Language, Bulgarian Academy of Sciences. Retrieved from https://aclanthology.org/2024.clib-1.7

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & **Manning, C. D.** (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations.* Retrieved from https://nlp.stanford.edu/pubs/qi2020stanza.pdf

Talamo, L., Steuer, J., & **Verkerk, A.** (submitted). They saw it, onu, tā, coming: An information theoretic study of cross-linguistic variation in personal pronouns. *Linguistics*.

Talamo, L., & **Verkerk, A.** (2022). A new methodology for an old problem: A corpus-based typology of adnominal word order in European languages. *Italian Journal of Linguistics*, *34*(2), 171–226.

Toska, M., Nivre, J., & **Zeman, D.** (2020). Universal Dependencies for Albanian. In: M.-C. de Marneffe, M. de Lhoneux, J. Nivre, & S. Schuster (Eds.), *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)* (pp. 178–188). Barcelona, Spain (Online): Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.udw-1.20

Verkerk, A., & **Talamo, L.** (2024). mini-CIEP + : A Shareable Parallel Corpus of Prose. In: P. Zweigenbaum, & R. Rapp, & S. Sharoff (Eds.), *Proceedings of the 17th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC-COLING 2024* (pp. 135–143). Torino, Italia: ELRA and ICCL. Retrieved from https://aclanthology.org/2024.bucc-1.15

TO CITE THIS ARTICLE:
Talamo, L. (2025). Introducing STAF: The Saarbrücken Treebank of Albanian Fiction. *Journal of Open Humanities Data*, 11: 3, pp. 1–6. DOI: https://doi.org/10.5334/johd.285

Submitted: 20 November 2024
Accepted: 28 December 2024
Published: 23 January 2025

COPYRIGHT:
© 2025 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/4.0/.

Journal of Open Humanities Data is a peer-reviewed open access journal published by Ubiquity Press.