# *Thamizhi*UDp: A Dependency Parser for Tamil

**Kengatharaiyer Sarveswaran**
University of Moratuwa / Sri Lanka
sarvesk@uom.lk

**Gihan Dias**
University of Moratuwa / Sri Lanka
gihan@uom.lk

## Abstract

This paper describes how we developed a neural-based dependency parser, namely *Thamizhi*UDp, which provides a complete pipeline for the dependency parsing of the Tamil language text using Universal Dependency formalism. We have considered the phases of the dependency parsing pipeline and identified tools and resources in each of these phases to improve the accuracy and to tackle data scarcity. *Thamizhi*UDp uses Stanza for tokenisation and lemmatisation, *Thamizhi*POSt and *Thamizhi*Morph for generating Part of Speech (POS) and Morphological annotations, and uuparser with multilingual training for dependency parsing. *Thamizhi*POSt is our POS tagger, which is based on the Stanza, trained with Amrita POS-tagged corpus. It is the current state-of-the-art in Tamil POS tagging with an F1 score of 93.27. Our morphological analyzer, *Thamizhi*Morph is a rule-based system with a very good coverage of Tamil. Our dependency parser *Thamizhi*UDp was trained using multilingual data. It shows a Labelled Assigned Score (LAS) of 62.39, 4 points higher than the current best achieved for Tamil dependency parsing. Therefore, we show that breaking up the dependency parsing pipeline to accommodate existing tools and resources is a viable approach for low-resource languages.

## 1 Introduction

Applying neural-based approaches to Tamil, like other Indic languages, is challenging due to a lack of quality data (Bhattacharyya et al., 2019), and the language's structure (Sarveswaran and Butt, 2019; Butt, Miriam, Rajamathangi, S. and Sarveswaran, K., 2020). Although there is a large volume of electronic unstructured/partially-structured text available on the Internet, not many language processing tools are publicly available for even fundamental tasks like part of speech (POS) tagging or parsing. Nowadays, neural-based approaches are the state of the art for most natural language processing tasks. These approaches require a significant amount of quality data for training and evaluation. On the other hand techniques like transfer learning, and multilingual learning may be used to overcome data scarcity. This paper discusses how we developed a neural-based dependency parser for Tamil with the aid of data orchestration and multilingual training.

## 2 Background and Motivation

Tamil is a Southern Dravidian language spoken by more than 80 million people around the world. However, it still lacks enough tools and quality annotated data to build good Natural Language Processing (NLP) applications.

### 2.1 Universal Dependency Treebank

Treebanks are a collection of texts with various levels of annotations, including Part of Speech (POS) and morpho-syntactic annotations. There are different formalisms used to mark syntactic annotations (Marcus et al., 1993; Böhmová et al., 2003; Nivre et al., 2016; Kaplan and Bresnan, 1982). Among the available formalisms, the dependency grammar formalism is useful for languages like Tamil which are morphologically rich, and whose word order is relatively variable and less bound (Bharati et al., 2009).

The Universal Dependency formalism (Nivre et al., 2016) is nowadays used widely to create Universal Dependency Treebanks (UD) with annotations. The current release of UDv2.7 has 183 annotated treebanks of various sizes from 104 languages (Zeman et al., 2020). UD captures information such as Parts of Speech (POS), morphological features, and syntactic relations in the form of dependencies. All these annotations are defined with multilingual language processing in mind, and the



*Proceedings of the 17th International Conference on Natural Language Processing*, pages 200–207
Patna, India, December 18 - 21, 2020. 

present format used to specify the annotation is called CoNLL-U format.[1]

There are only three Indic languages, namely, Hindi, Urdu, and Sanskrit that have relatively large datasets 375K, 138K, and 28K tokens, respectively in UDv2.7. All the other six Indic languages, including Tamil and Telugu, have less than 12K tokens in UDv2.7.

### 2.2 Tamil Universal Dependency Treebanks

Tamil has been included in UD treebank releases since 2015. Initially it was populated from the Prague Style Tamil treebank by (Ramasamy and Žabokrtský, 2012), and since then the dataset has been part of the UD without much alterations or corrections. Tamil TTB in UDv2.6 has some inaccuracies, and inconsistencies. For instance, numbers are marked as NUM and ADJ, while only the former tag is correct. The first author of this paper has corrected some of these issues and made it available in UDv2.7. However, there are still more issues that need to be solved. Tamil TTB in UDv2.7 has altogether 600 sentences for training, development and testing. In (Zeman et al., 2020), there is another Tamil treebank with 536 sentences, namely MWTT, which has been newly added. MWTT is based on the Enhanced Universal Dependency[2] annotation, where complex concepts like elision, relative clauses, propagation of conjuncts, raising and control constructions, and extended case marking are captured. Therefore, there are slight variations in TTB and MWTT. Further, MWTT has very short sentences, while TTB has relatively very longer ones. In this paper, we have mainly used and discussed Tamil TTB.

### 2.3 Dependency parsers

A Dependency parser is a type of syntactic parser which is useful to elicit lexical, morphological, and syntactic features, and the inter-connections of tokens in a given sentence. Linguistically, this would be useful for syntactic analyses, and comparative studies. Computationally, this is a key resource for natural language understanding (Dozat and Manning, 2018). Different approaches are employed when developing dependency parsers. However, neural-based parsers are the latest state of the art.

There are several off-the-shelf neural-based parsers available that are built around Universal Dependency Treebanks (UD), including Stanza (Qi et al., 2020), and uuparser (de Lhoneux et al., 2017) and its derivatives. Both of these are open source tools. Stanza is a Python NLP library which includes a multilingual neural NLP pipeline for dependency parsing using Universal Dependency formalism. uuparser is a tool developed specifically for UD parsing. These neural-based tools need large amount of quality data on which to be trained.

On the other hand, several approaches are being used to overcome the issue with data scarcity, including multilingual training. There is an attempt to create a multilingual parsing for several low-resource languages, and it is reported that multilingual training significantly improves the parsing accuracy of low-resource languages (Smith et al., 2018).


Figure 1: Phases of the Parsing Pipeline
Image source: https://stanfordnlp.github.io/stanza

## 3 *Thamizhi*UDp

By considering all available resources and approaches as outlined in section 2, we decided to develop a Universal Dependency parser (UDp) for Tamil called *Thamizhi*UDp using existing open source tools, namely Stanza, *Thamizhi*Morph, and uuparser. However, since we do not have enough data to train a neural-based parser end-to-end, we have broken up the pipeline to different phases. We have then orchestrated data from different sources for each of these phases, and used different tools in different phases, as shown in Table 1. The following sub-sections discuss each of the stages of the pipeline, and how we went about developing them.

Our dependency parsing pipeline has several stages as shown in Figure 1. As mentioned, we used different datasets and tools, as shown in Table 1, in the different stages of the pipeline. Currently,

[1] https://universaldependencies.org/format.html
[2] https://universaldependencies.org/u/overview/enhanced-syntax.html

| Step | Tool | Dataset |
|---|---|---|
| Tokenisation | Stanza | Tamil UDT |
| Multi-word tokeniser | Stanza | Tamil UDT |
| Lemmatisation | Stanza | Tamil UDT |
| POS tagging | *Thamizhi*POSt | Amrita Data |
| Morphological tagging | *Thamizhi*Morph | Rule-based |
| Dependency parsing | uuparser | UDT of various languages |

Table 1: *Thamizhi*UDp process pipeline

Stanza does not have support for multilingual training. Therefore, for dependency parsing, we used uuparser with the multilingual training. Each of the phases within the pipeline is explained in the following respective sub-sections.

### 3.1 Tokenisation

First, the given texts have been Unicode normalised, and then tokenised, and broken up in to sentences. We developed a script[3] to do Unicode normalisation. Because of different input methods or other reasons, at times the same surface form of a character has been stored using different Unicode sequences. Therefore, this needed to be normalised, otherwise, a computer would consider them as different characters. Once this normalisation was done, we moved on to tokenisation. To do this, we trained Stanza with the texts available in TTB. During this phase, punctuations were separated from words, and the given texts were broken in to sentences.

### 3.2 Multi-word tokenisation using Stanza

After the initial tokenisation, syntactically compound words or multi-word tokens were broken into syntactic units as proposed by the UD guidelines,[4] so that syntactic dependencies can be marked precisely. Syntactically compound constructions are common in Tamil. For instance, words with *-um* clitic will be tokenised, like *naanum* =⟩ *naan+um* 'I+and', so that coordinating conjunctive dependency can be shown easily. In the current TTB UDv2.7, there are 520 instances of multi-words found among 400 sentences in the training set. We used this TTB training set to train our multi-word tokeniser using Stanza. However, multi-word tokenisations are not properly divided in TTB. We are in the process of improving this multi-word tokeniser with the use of more data.

### 3.3 Lemmatisation using Stanza

UD Treebanks also have lemmas marked in their CoNLL-U format annotation This is useful for language processing applications, such as a Machine Translator. We trained Stanza using the TTB UDv2.6 to do lemmatisation. However, the current TTB has several inaccuracies in identifying lemmas, specifically due to improper multi-word tokenisation. Since a lemma is identified for multi-word tokenised words, multi-word tokenisation has an effect on lemmatisation. Since we are still in the process of improving our multi-word tokeniser, lemmatisation will also be improved in the future.

### 3.4 POS tagging using *Thamizhi*POSt

Part of Speech (POS) tagging is an important phase in the parsing process where each word in a sentence is assigned with its POS tag (or lexical category) information. Several attempts have been made to define POS tagsets for Tamil, based on different theories, and level of granularity; (Sarveswaran and Mahesan, 2014) gives an account of different tagsets. Among these, Amrita (Anand Kumar et al., 2010) and BIS[5] are two popular tagsets. In addition to tagsets, Amrita and BIS POS tagged data are also available. The corpus[6] which is tagged using BIS tagset is taken from a historical novel, while the corpus tagged using Amrita is taken from news websites. Further more we found that Amrita's data is cleaner and there is more consistency when it comes to POS tagging. We also harmonised the tags found in the BIS, Amrita, and UPOS[7] tagsets.

Though there have been several attempts to develop a POS tagger for Tamil, there are not available, or have not given convincing results. Moreover, only a few neural-based approaches for Tamil POS tagging have been developed. Therefore, we decided to develop a POS tagger, namely *Thamizhi*POSt, using Stanza, and to publish it as an open source tool. We used the corpus tagged using Amrita's POS tagged corpus to train this tagger. The development process is outlined briefly below.

[3] https://github.com/sarves/thamizhi-validator
[4] https://universaldependencies.org/u/overview/tokenization.html
[5] http://www.tdil-dc.in/tdildcMain/articles/134692DraftPOSTagstandard.pdf
[6] http://www.au-kbc.org/nlp/corpusrelease.html
[7] https://universaldependencies.org/u/pos/all.html

| Neural-based POS taggers | F1 Score |
|---|---|
| PVS and Karthik (2007) | 87.0* |
| Mokanarangan et al. (2016) | 87.4 |
| Qi et al. (2020) | 82.6** |
| ***Thamizhi*POSt** | **93.27** |

Table 2: Scores of neural-based POS taggers for Tamil POS tagging
*github.com/avineshpvs/indic_tagger
**stanfordnlp.github.io/stanza/performance.html

First we mapped Amrita's 32 POS tags to Universal POS (UPOS); see Table 4 in Appendix A for the mapping of Amrita-UPOS mapping. In doing so, we converted the annotations from Amrita POS tags to UPOS tags. We then divided Amrita POS tagged corpus of 17K sentences in to 11K, 5K and 1K sentences for training, development, and testing, respectively. Thereafter, we converted these datasets in CoNLL-U format so that it could then be fed to Stanza. Following that we trained and evaluated *Thamizhi*POSt, which is a Stanza instance that has been trained on Amrita's data. During the training, we also used fastText model (Bojanowski et al., 2016) to capture the context in POS tagging, as specified in Stanza.

We also evaluated *Thamizhi*POSt using Tamil UDv2.6 test data. The F1 score of the evaluation was 93.27, which is higher than the results reported for existing neural-based POS taggers, as shown Table 2.

### 3.5 Morphological tagging

We used an open source morphological analyser called (Sarveswaran et al., 2019, 2018),[8] which we developed as part of our project on computational grammar for Tamil, to generate morphological features according to the UD specification.[9] Since we have developed this rule-based analyser for grammar development purposes, this gives us a very detailed analysis for each given word. We used *Thamizhi*Morph in the process. At this stage, we fed the tokenised, lemmatised and POS tagged data in the CoNLL-U format to *Thamizhi*Morph to do the morphological analyses.

As a morphological analyser, *Thamizhi*Morph gives us all the possible morphological analyses for a given word. In addition to the morphological analysis, it also gives us the POS tag information, and lemma information. When the lemma of a given surface form in not found in the *Thamizhi*Morph lexicon, it uses a rule-based guesser to predict the lemma; sometimes this fails too, especially when there is a foreign word.

However, for our parsing purpose, we wanted to get the single correct morphological analysis based on the context. This was challenging. To tackle this challenge, we used a disambiguation process to generate a single analysis. However, we still failed at times, since we especially get multiple analyses because of the way some people write. When this was the case, we manually picked the correct analysis, even after our disambiguation process. We are now in the process of training a Stanza based morphological analyser using the data generated by *Thamizhi*Morph. We hope this will improve the robustness, especially when there are out of vocabulary tokens.

### 3.6 Dependency parsing

When we looked for a Dependency parser, we found that none existed that were specifically trained for Tamil. For TTB test data, in their default configurations the off-the-shelf Stanza and uuparser give the Labelled Assigned Score (LAS) of 57.64 and 55.76, respectively.

We wanted to improve the accuracy, however, we could not find any datasets with dependency annotations, other than TTB UDv2.6 at the time of development. To overcome this data scarcity, we tried multilingual training for Tamil along with Hindi HDTB,[10] Turkish,[11] Arabic,[12] and Telugu,[13] which we found would be relevant, available in UDv2.6. We did this multilingual training using uuparser. The experiment gave us some good results, when we compared this with what was reported by Stanza or uuparser as shown in Table 3.

As in Table 3, we got a LAS of 62.39 when training with Hindi HDTB UDv2.6, but, surprisingly, not when training with Telugu, which is also a Dravidian language like Tamil. We trained the tagger with the whole Telugu, and Hindi treebanks along with Tamil. However, the score was lesser than what we got when we trained it with Hindi

[8] https://github.com/sarves/thamizhi-morph
[9] https://universaldependencies.org/u/feat/all.html
[10] https://github.com/UniversalDependencies/UD_Hindi-HDTB/tree/master
[11] https://github.com/UniversalDependencies/UD_Turkish-IMST/tree/master
[12] https://github.com/UniversalDependencies/UD_Arabic-PADT/tree/master
[13] https://github.com/UniversalDependencies/UD_Telugu-MTG/tree/master

| Languages (# of sent.) | Accuracy(LAS) |
|---|---|
| with Telugu (100) | 58.91 |
| with Telugu ( 1050) | 59.22 |
| **with Hindi (1600)** | **62.39** |
| with Telugu (100) and Arabic (100) | 58.04 |
| with Telugu (100) and Turkish (100) | 58.43 |
| with Telugu (100) and Hindi (100) | 59.07 |

Table 3: LAS of Multilingual parsing

data. For all these experiments, we used the Tamil testing set available in TTB UDv2.6.

## 4 Discussion

Tamil TTB has not undergone any major revisions or corrections since its initial release. It has several issues, in POS tagging, multi-word tokenisation, and dependency tagging. Altogether we only have 600 sentences for training, development, and testing; some of these sentences are very long. All these made the training of a UD parser a difficult task. We tried to overcome some of these issues using other data, and tools available online. However, we still depend on this dataset for some part of the training, such as for dependency parsing.

Only one treebank, Telugu MTG UDv2.6, which is the closest to Tamil in terms of linguistic structures, is available as of today. We observed that Telugu UDv2.6 is small in size. That only has around 1050 sentences compared to Hindi HDTB UDv2.6. Moreover, Telugu has very short sentences without any morphological feature information. On the other hand, some sentences in TTB in UDv2.6 has up to 40 tokens. Because of all these varied factors we could not achieve much improvement when use Telugu MTG UDv2.6 in multilingual training. However, Hindi, which belongs to a different language family, showed better performance when used for Multilingual training. We have additionally noticed that the accuracy of the dependency parsing also improved when we increased the Hindi data size during the training.

Another challenge we have faced was finding quality test data or benchmark datasets for evaluation. In the current practice, everyone tests their tools using their own dataset to evaluate. Therefore, it is always a challenging task to reproduce or compare results. In our case, for dependency parsing, we used the UD test data. However, it is not a clean and error free dataset for evaluation. For this reason, we have now started working on a Tamil dependency treebank which can soon be used as an evaluation dataset.

We used our personal computers without any Graphical Processing Units (GPU) to carry out all these experiments. However, high performance computing resources will save time, and we might need to go for such resources when we increase the size of datasets.

## 5 Conclusion

e have implemented a Universal Dependency parser for Tamil, *Thamizhi*UDp, which annotates a Tamil sentence with POS, Lemma, Morphology, and Dependency information in CoNLL-U format. We have developed a parsing pipeline using several open source tools and datasets to overcome data scarcity. We have also used multilingual training to overcome the scarecity of dependency annotated data. *Thamizhi*POSt, a POS tagger for Tamil, has been implemented using Stanza and the Amrita POS tagged dataset. *Thamizhi*POSt outperforms existing neural-based POS taggers, and gives an F1 score of 93.27. Further, we obtained the best accuracy of LAS 62.39 for dependency parsing in a multilingual training setting with Hindi HDTB, using uuparser. More importantly, we have made our tools *Thamizhi*POSt,[14,15] *Thamizhi*Morph,[16,17] and *Thamizhi*UDp[18,19] along with relevant models, datasets, and scripts available open source for others to use and extend upon.

## Acknowledgements

We would like express our appreciation to Maris Camilleri from the University of Essex for her support in language editing, and three anonymous reviewers for their valuable comments and inputs to improve this menu script.

This research was supported by the Accelerating Higher Education Expansion and Development (AHEAD) Operation of the Ministry of Higher Education, Sri Lanka funded by the World Bank.

[14] http://nlp-tools.uom.lk/thamizhi-pos
[15] https://github.com/sarves/thamizhi-pos/
[16] http://nlp-tools.uom.lk/thamizhi-morph
[17] https://github.com/sarves/thamizhi-morph/
[18] http://nlp-tools.uom.lk/thamizhi-udp
[19] https://github.com/sarves/thamizhi-udp/

## References


M Anand Kumar, V Dhanalakshmi, KP Soman, and S Rajendran. 2010. A sequence labeling approach to morphological analyzer for Tamil language. *International Journal on Computer Science and Engineering (IJCSE)*, 2(06):1944–195.

Akshar Bharati, Mridul Gupta, Vineet Yadav, Karthik Gali, and Dipti Misra Sharma. 2009. Simple parser for Indian languages in a dependency framework. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 162–165.

Pushpak Bhattacharyya, Hema Murthy, Surangika Ranathunga, and Ranjiva Munasinghe. 2019. Indic language computing. *Communications of the ACM*, 62(11):70–75.

Alena Böhmová, Jan Hajič, Eva Hajičová, and Barbora Hladká. 2003. The Prague Dependency Treebank. In *Treebanks*, pages 103–127. Springer.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Butt, Miriam, Rajamathangi, S. and Sarveswaran, K. 2020. Mixed Categories in Tamil via Complex Categories. In *Proceedings of the LFG20 Conference*, Stanford. CSLI Publications.

Timothy Dozat and Christopher D. Manning. 2018. Simpler but more accurate semantic dependency parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 484–490, Melbourne, Australia. Association for Computational Linguistics.

Ron Kaplan and Joan Bresnan. 1982. Lexical functional grammar: a formal system for grammatical representation. The Mental Representation of Grammatical Relations. *J. Bresnan. Cambridge, MA: MIT Press*, pages 173–281.

Miryam de Lhoneux, Yan Shao, Ali Basirat, Eliyahu Kiperwasser, Sara Stymne, Yoav Goldberg, and Joakim Nivre. 2017. From raw text to universal dependencies-look, no tags! In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 207–217.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2):313–330.

T Mokanarangan, T Pranavan, U Megala, N Nilusija, Gihan Dias, Sanath Jayasena, and Surangika Ranathunga. 2016. Tamil morphological analyzer using support vector machines. In *International Conference on Applications of Natural Language to Information Systems*, pages 15–23. Springer.

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

Avinesh PVS and G Karthik. 2007. Part-of-speech tagging and chunking using conditional random fields and transformation based learning. *Shallow Parsing for South Asian Languages*, 21:21–24.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Loganathan Ramasamy and Zdeněk Žabokrtský. 2012. Prague dependency style treebank for Tamil. In *Proceedings of Eighth International Conference on Language Resources and Evaluation (LREC 2012)*, pages 1888–1894, İstanbul, Turkey.

K Sarveswaran, Gihan Dias, and Miriam Butt. 2018. ThamizhiFST: A Morphological Analyser and Generator for Tamil Verbs. In *2018 3rd International Conference on Information Technology Research (ICITR)*, pages 1–6. IEEE.

K Sarveswaran, Gihan Dias, and Miriam Butt. 2019. Using meta-morph rules to develop morphological analysers: A case study concerning Tamil. In *Proceedings of the 14th International Conference on Finite-State Methods and Natural Language Processing*, pages 76–86, Dresden, Germany. Association for Computational Linguistics.

K Sarveswaran and S Mahesan. 2014. Hierarchical Tag-set for Rule-based Processing of Tamil Language. *International Journal of Multidisciplinary Studies (IJMS)*, 1(2):67–74.

Kengatharaiyer Sarveswaran and Miriam Butt. 2019. Computational Challenges with Tamil Complex Predicates. In *Proceedings of the LFG19 Conference, Australian National University*, pages 272–292, Stanford. CSLI Publications.

Aaron Smith, Bernd Bohnet, Miryam de Lhoneux, Joakim Nivre, Yan Shao, and Sara Stymne. 2018. 82 treebanks, 34 models: Universal Dependency parsing with multi-treebank models. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 113–123, Brussels, Belgium. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Gabrielė Aleksandravičiūtė,

Ika Alfina, Lene Antonsen, Katya Aplonova, Angelina Aquino, Carolina Aragon, Maria Jesus Aranzabe, Hórunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Kepa Bengoetxea, Gözde Berk, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Kristín Bjarnadóttir, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Ethan Chi, Yongseok Cho, Jinho Choi, Jayeol Chun, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Arawinda Dinakaramani, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Marília Fernanda, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Fabrício Ferraz Gerardi, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Hinrik Hafsteinsson, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Sam Hardwick, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Eva Huber, Jena Hwang, Takumi Ikeda, Anton Karl Ingason, Radu Ion, Elena Irimia, Ọlájídé Ishola, Tomáš Jelínek, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Sarveswaran K, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Sookyoung Kwak, Veronika Laippala, Lucia Lam, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phng Lê H`ông, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Cheuk Ying Li, Josie Li, Keying Li, Yuan Li, KyungTae Lim, Krister Lindén, Nikola Ljubešić, Olga Loginova, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Karina Mischenkova, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Keiko Sophie Mori, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Lng Nguy˜ên Thị, Huy`ên Nguy˜ên Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Alireza Nourian, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ̀ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Arzucan Özgür, Balkız Öztürk Başaran, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Eiríkur Rögnvaldsson, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Olga Rudina, Jack Rueter, Kristján Rúnarsson, Shoval Sadde, Pegah Safari, Benoît Sagot, Aleksi Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Kevin Scannell, Salvatore Scarlata, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Muh Shohibussirri, Dmitry Sichinava, Einar Freyr Sigursson, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Steinþór Steingrímsson, Antonio Stella, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki,

Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Samson Tella, Isabelle Tellier, Guillaume Thomas, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika Vincze, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, Hanzhi Zhu, and Zhuravleva. 2020. Universal Dependencies 2.7. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

## Appendix A: Harmonisation of Amrita and UPOS tagsets

| **Amrita** | **UPOS** | **Amrita** | **UPOS** |
|---|---|---|---|
| NN | NOUN | CVB | CCONJ |
| NNC | NOUN | PPO | ADP |
| NNP | PROPN | CNJ | CCONJ |
| NNPC | PROPN | DET | DET |
| ORD | NUM | COM | CCONJ |
| CRD | NUM | EMP | PART |
| PRP | PRON | ECH | PART |
| PRIN | PRON | RDW | ADP |
| ADJ | ADJ | QW | VERB |
| ADV | ADV | QM | PUNCT |
| VNAJ | VERB | INT | ADJ |
| VNAV | VERB | NNQ | NUM |
| VINT | VERB | QTF | NUM |
| VBG | NOUN | COMM | PUNCT |
| VF | VERB | DOT | PUNCT |
| VAX | VAUX | | |

Table 4: Harmonisation of Amrita and UPOS tagsets