Title	Overview of part-of-speech taggers and lemmatizers
Version	1.0
Author(s)	Darja Fišer, Jakob Lenardič
Date	30-03-2020
Status	For distribution
Distribution	BoD, NCF, UI
ID	CE-2020-1641

Table of contents

1. Introduction

In this report, we present an overview of language tools dedicated to part-of-speech tagging, and/or lemmatization. Part-of-speech tagging is the automatic text annotation process in which words or tokens are assigned part of speech tags, which typically correspond to the main syntactic categories in a language (e.g., noun, verb, adverb) and often to subtypes of a particular syntactic category which are distinguished by morphosyntactic features (e.g.,

pronouns are assigned different part-of-speech tags based on their case).[1] Lemmatization is the process by which inflected forms of a lexeme are grouped together under a base dictionary form. Part-of-speech tagging and lemmatization are crucial steps of linguistic pre-processing.

The overview was conducted in three steps:

(i) By overviewing the Virtual Language Observatory (VLO) or the CLARIN Language Switchboard, we prepared a preliminary survey of taggers and lemmatizers in a Google Docs document. The VLO was searched with the keywords *part-of-speech tagger*, *part of speech tagger*, *lemmatizer*, and *lemmatiser*, as well as with the terms *lemmatization, lemmatization*, and *tagger*, subsequently narrowing the results down by the Resource Type facet to values related to software.

(ii) Afterwards, we asked national CLARIN UI representatives to provide information on part-of-speech tagging and lemmatization from their own countries in the Google Docs survey.[2]

(iii) Finally, we added a few additional tools listed in the VLO or the CLARIN Language Switchboard.

In this way we collected a set of 65 tools dedicated to part-of-speech tagging and/or lemmatization. Our primary aim of this survey was to evaluate the presentation of their availability and metadata (primarily language, functionality and licence).

2. Taggers and lemmatizers in the CLARIN infrastructure

There are 65 tools for PoS/MSD-tagging and/or lemmatization in the CLARIN infrastructure. The tools are described in Sections 2.1 and 2.2 with the following metadata:

(i) Functionality
(ii) Licence
(iii) Availability
(iv) Input and output formats
(v) Publication

Section 3 summarizes the metadata.

[1] In the report, we use the acronyms PoS, which stands for *part-of-speech*, and MSD, which stands for *morphosyntactic description*. MSD tags denote fine-grained feature-structure based PoS tags which are used to account for rich inflectional paradigms like those in Slavic languages. However, since part of speech tags are often differentiated from one another not just by their syntactic category but also by morphological information even in the case of morphologically poor languages like English, we group tools described as MSD-taggers together with tools described as PoS-taggers in terms of functionality (see Section 3.3.2).

[2] We would like to thank all the UI representatives and National Coordinators who have participated in the survey.

2.1. Tools for tagging/lemmatizing within a single language

Table 1 lists 50 tools for PoS/MSD-tagging and/or lemmatization within a single language.

Table 1: PoS/MSD-taggers and lemmatizers for a single language, sorted by language and corpus name

Tool	Languages	Description
Afrikaans TnT-Tagger **Functionality:** PoS	Afrikaans	This tool is based on the TnT tagger (Brants 2000). The tagset used by the tool was especially designed for Afrikaans and consists of 139 PoS-tags. **Input:** plain text **Output:** plain text **CLARIN Centre:** SADiLaR
NCHLT Afrikaans Lemmatiser **Functionality:** lemma **Licence:** CC-BY 2.5 South Africa Licence	Afrikaans	This tool is a lemmatizer for Afrikaans developed during the NCHLT Text project (Barnard et al. 2014). **Availability:** download **CLARIN Centre:** SADiLaR
Assamese POS Tagger **Functionality:** PoS	Assamese	This tool is a CRF++ based PoS-tagger. **CLARIN Centre:** CLARIN-PL
NCHLT isiNdebele Lemmatiser **Functionality:** lemma **Licence:** CC-BY 2.5 South Africa Licence	Bantu	This tool is a lemmatizer for Ndebele Bantu language developed during the NCHLT Text project (Barnard et al. 2014). **Availability:** download **CLARIN Centre:** SADiLaR
NCHLT isiXhosa Lemmatiser **Functionality:** lemma **Licence:** CC-BY 2.5 South Africa Licence	Bantu	This tool is a lemmatizer for the Xhosa Bantu language developed during the NCHLT Text project (Barnard et al. 2014). **Availability:** download **CLARIN Centre:** SADiLaR
NCHLT isiZulu Lemmatiser **Functionality:** lemma **Licence:** CC-BY 2.5 South Africa Licence	Bantu	This tool is a lemmatizer for the Zulu Bantu language developed during the NCHLT Text project (Barnard et al. 2014). **Availability:** download **CLARIN Centre:** SADiLaR
NCHLT Sepedi Lemmatiser **Functionality:** lemma **Licence:** CC-BY 2.5 South Africa Licence	Bantu	This tool is a lemmatizer for the Sepedi (Northern Sotho) Bantu language developed during the NCHLT Text project (Barnard et al. 2014). **Availability:** download **CLARIN Centre:** SADiLaR

NCHLT Sesotho Lemmatiser **Functionality:** lemma **Licence:** CC-BY 2.5 South Africa Licence	Bantu	This tool is a lemmatizer for the Sesotho Bantu language developed during the NCHLT Text project (Barnard et al. 2014). **Availability:** download **CLARIN Centre:** SADiLaR
NCHLT Setswana Lemmatiser **Functionality:** lemma **Licence:** CC-BY 2.5 South Africa Licence	Bantu	This tool is a lemmatizer for the Tswana Bantu language developed during the NCHLT Text project (Barnard et al. 2014). **Availability:** download **CLARIN Centre:** SADiLaR
NCHLT Siswati Lemmatiser **Functionality:** lemma **Licence:** CC-BY 2.5 South Africa Licence	Bantu	This tool is a lemmatizer for the Swazi Bantu language developed during the NCHLT Text project (Barnard et al. 2014). **Availability:** download **CLARIN Centre:** SADiLaR
NCHLT Tshivenda Lemmatiser **Functionality:** lemma **Licence:** CC-BY 2.5 South Africa Licence	Bantu	This tool is a lemmatizer for the Venda Bantu language developed during the NCHLT Text project (Barnard et al. 2014). **Availability:** download **CLARIN Centre:** SADiLaR
NCHLT Xitsonga Lemmatiser **Functionality:** lemma **Licence:** CC-BY 2.5 South Africa Licence	Bantu	This tool is a lemmatizer of the Tsonga Bantu language developed during the NCHLT Text project (Barnard et al. 2014). **Availability:** download **CLARIN Centre:** SADiLaR
Sepedi Part of Speech Tagger **Functionality:** PoS	Bantu	This tool is based on Helmut Schmidt stochastic tagger (see Schmid 1994) supported by additional noun and verb guessing modules and a tokenizer. **CLARIN Centre:** SADiLaR
Corpus.by Lemmatizer **Functionality:** lemma	Belarussian	This tool is part of the corpus.by platform. **Availability:** web service **CLARIN Centre:** CLARIN Knowledge Centre for Belarusian text and speech processing **Input:** plain text **Output:** plain text
CLaRK **Functionality:** sentence splitting, PoS,	Bulgarian	This tool is an XML-based software system for corpora development implemented in JAVA. The main aim behind the design of the system is the minimization of human intervention

lemma, syntactic parsing		during the creation of language resources. CLaRK includes BTB-Pipe, which is a language pipeline for Bulgarian that comprises the following modules: sentence splitting, MSD-tagging, lemmatization, dependency parsing. **Availability:** dowload **Input:** XML **Output:** XML **CLARIN Centre:** ClaDA-BG **Related publication:** Simov et al. (2001)
HMM tagger **Functionality:** MSD **Licence:** GNU General Public Licence, version 2	Czech	This tool uses Hidden Markov Models and is an implementation of the UFAL tagger. **Availability:** download **CLARIN Centre:** LINDAT
Frog **Functionality:** PoS, MSD, lemma, NE, phrase chunks, dependency relations with head words **Licence:** GNU General Public Licence	Dutch	This tool is an integration of memory-based NLP modules developed for Dutch. All NLP modules are based on TiMBL, the Tilburg memory-based learning software package. Where possible, Frog makes use of multi-processor support to run subtasks in parallel. **Availability:** download **Output:** FoLiA XML **CLARIN Centre:** CLARIAH-NL **Related publication:** Van den Bosch et al. (2007)
INL Labs tagger/lemmatizer tools **Functionality:** PoS, lemma **Licence:** CLARIN PUB	Dutch	This tool employs a PoS tagger that is trained on the "Letters as loot" historical corpus and a lemmatizer that is trained on the INL historical lexicon. **Availability:** web application **CLARIN Centre:** CLARIAH-NL **Input:** plain text, TEI, epub, html, docx, alto **Output:** styled, XML
Tadpole **Functionality:** PoS/MSD, lemma, syntactic parsing	Dutch	An integrated tokenizer, tagger-lemmatizer, morphological analyzer, and dependency parser for Dutch. **Availability:** broken link **CLARIN Centre:** LINDAT/CLARIAH-NL
MorphAdorner Lemmatizer	English	This tool is implemented in WebLicht and is derived from the MorphAdorner morphological analyser.

Functionality: lemma		**Availability:** WebLicht **Input:** TCF, XML **CLARIN Centre:** CLARIN-D
OpenNLP Part-of-Speech Tagger (English) **Functionality:** PoS **Licence:** Apache Licence 2.0 (restricted)	English	This tool is based on the Apache OpenNLP library, which is a perception and maximum entropy-based machine learning toolkit for the processing of natural language text. **Availability:** web application **Input:** application/xml **Output:** application/xml **CLARIN Centre:** CLARIN:EL
Stanford Dependency Parser **Functionality:** PoS, syntactic parsing	English	This tool is a WebLicht implementation of the Stanford Parser. **Availability:** WebLicht **Input:** plain text, pdf, rtf, XML **Output:** plain text, pdf, rtf, XML **CLARIN Centre:** CLARIN-D **Related publication:** Hinrichs et al. (2010)
EstNLTK **Functionality:** MSD, NER **Licence:** Available - Unrestricted Use	Estonian	This tool provides common natural language processing functionality such as morphological analysis and named entity recognition for the Estonian language. Web API documentation is available here. **Availability:** download **Input:** plain text **Output:** plain text **CLARIN Centre:** CELR **Related publication:** Orasmaa et al. (2016)
Vabamorf open source morphology tagger for Estonian **Functionality:** PoS, MSD, lemma **Licence:** Available - Unrestricted Use	Estonian	This tool performs various tasks of morphological analysis, including morphological disambiguation and synthesis. **Availability:** download, web application **Input:** plain text **Output:** plain text **CLARIN Centre:** CELR **Related publication:** Kaalep (2015)
FinTag **Functionality:** PoS, lemma, NER **Licence:** GPL	Finnish	This toolchain provides finnish-postag, a part-of-speech and morphology tagger for Finnish, and finnish-nertag, a named entity recogniser for Finnish. Both tools take running text from

		standard input and produce tabular output (one token per line) to standard output. **Availability:** download, web application **Input:** plain text, pdf, doc, scv, epub, html, odt, xls **Output:** TSV **CLARIN Centre:** FIN-CLARIN
OpenNLP Part-of-Speech Tagger (German) **Functionality:** PoS **Licence:** Apache Licence 2.0 (restricted)	German	This tool is based on the Apache OpenNLP library, which is a perception and maximum entropy–based machine learning toolkit for the processing of natural language text. **Availability:** web application **Input:** application/xml **Output:** application/xml **CLARIN Centre:** CLARIN:EL
WebLicht Part-of-Speech Tagger **Functionality:** PoS, lemma	German	This tool is a PoS tagger and lemmatizer implemented in WebLicht. **Availability:** web application, WebLicht **Input:** TCF, XML **CLARIN Centre:** CLARIN-D
SepVerb Lemmatizer **Functionality:** lemma	German	This tool is based on the Mate toolkit. **Availability:** WebLicht **Input:** TCF, XML **CLARIN Centre:** CLARIN-D
SMOR lemmatizer **Functionality:** PoS, lemma	German	This tool is implemented in WebLicht. **Availability:** broken link **Input:** TCF, XML **CLARIN Centre:** CLARIN-D
Stuttgart Dependency Parser **Functionality:** PoS, syntactic parsing	German	This tool is a Weblicht implementation of the Stuttgart parser. **Availability:** WebLicht **Input:** plain text, pdf, rtf, XML **Output:** plain text, pdf, rtf, XML **CLARIN Centre:** CLARIN-D **Related publication:** Hinrichs et al. (2010)
ILSP Feature-based multi-tiered POS Tagger **Functionality:** PoS	Greek	This tool is a FBT-based multitiered tagger. FBT is a variant of the well-known transformation based learning paradigm aiming at improving the quality of tagging highly inflective languages such as Greek.

Licence: terms of service (Restrictions: Academic - Non Commercial Use)		**Availability:** web application **Input:** Application/vnd.xmi+xml **Output:** Application/vnd.xmi+xml **CLARIN Centre:** CLARIN:EL **Related publication:** Papageorgiou et al. (2000)
hunpos **Functionality:** PoS **Licence:** New BSD License	Hungarian	This tool is an open source reimplementation of the TnT tagger (Brants 2000). **Availability:** download **CLARIN Centre:** LINDAT **Related publication:** Halácsy et al. (2007)
IceNLP Natural Language Processing toolkit **Functionality:** PoS, lemma, shallow syntactic parsing **Licence:** GNU General Public License, version 2	Icelandic	This tool is an open source NLP toolkit for analyzing and processing Icelandic text. The toolkit is implemented in Java. **Availability:** download, web application **Input:** plain text **Output:** plain text **CLARIN Centre:** CLARIN-IS **Related publication:** Loftsson and Rögnvaldsson (2007)
Freeling **Functionality:** PoS, lemma	Italian	This toolchain was developed in the PANACEA project and implements Freeling 2.1 libraries. **Availability:** web application **CLARIN Centre:** CLARIN-IT **Publication:** Padró et al. (2010)
NLP-PIPE **Functionality:** MSD, syntactic parsing, NER **Licence:** GNU General Public Licence 3	Latvian	This tool is a modular toolchain that allows researchers to combine multiple natural language processing tools in a unified framework. It provides the gluing code that is used to combine tools even if they are written in different programming languages and rely on conflicting library versions. It was created to make NLP technology more accessible to linguists, and to make new tool creation and integration easier to researchers and software developers. **Availability:** download, **CLARIN Centre:** CLARIN-LV **Related publication:** Znotins and Cirule (2018)
MLSS Tagger Web Service **Functionality:** PoS **Licence:** CLARIN ACA	Maltese	This tool is an implementation of the TnT tagger (Brants 2000). The model for Maltese was trained on manually tagged texts and has reached an accuracy of 96%. The tagset tailored to Maltese is available here.

		Availability: web application **CLARIN Centre:** PORTULAN
The Oslo-Bergen tagger **Functionality:** MSD, syntactic parsing **Licence:** GNU General public licence	Norwegian	This tool consists of three main modules: a pre-processor with a composition analyzer and multitagger, a grammar module for morphological and syntactic disambiguation (based on the constraint grammar paradigm) and a statistical module that removes the last residual morphological ambiguity (only for bookmarks). The tool is trained on the Norwegian wordbank. **Availability:** download **CLARIN Centre:** CLARINO **Related publication:** Johannessen et al. (2012)
Morfeusz 2 **Functionality:** MSD **Licence:** BSD 2 (public)	Polish	This tool is a dictionary-based morphological analyser and generator for Polish. This version of the program is decoupled from the dictionary. Two dictionaries of Polish developed within other projects are distributed with Morfeusz 2, namely SGJP and Polimorf. **Availability:** download, web application **Input:** various **Output:** various **CLARIN Centre:** CLARIN-PL **Related publication:** Woliński (2014)
MorphoDiTa-based tagger for Polish language **Functionality:** MSD **Licence:** GNU LGPL 3.0	Polish	This tool is based on the MorphoDiTa tagger, adapted to Polish. The tool employs the NKJP tagset. **Availability:** download **CLARIN Centre:** CLARIN-PL
Tagger SentiOne - version 2 **Functionality:** MSD **Licence:** GNU GPL3	Polish	This tool is the second version of tagger developed in the sentione project, adapted to UGC-processing. The tool has been enriched with some heuristics to improve its accuracy and a tokenizer. **Availability:** download **CLARIN Centre:** CLARIN-PL
Tagger WS **Functionality:** MSD, lemma	Polish	This tool uses the NKJP tagset and implements the Morfeusz SGJP dictionary. The service is based on WCRFT. **Availability:** web application

		Input: plain text, XML **Output:** plain text, XML **CLARIN Centre:** CLARIN-PL
TaKIPI **Functionality:** MSD	Polish	This tool assumes the morpho-syntactic description of the IPI PAN corpus tagset (Przepiórkowski 2005). **CLARIN Centre:** CLARIN-PL **Related publication:** Piasecki (2007)
WCRFT (Wrocław CRF Tagger) **Functionality:** MSD **Licence:** GNU LGPL 3.0 (info missing in the CLARIN-PL entry)	Polish	This tool combines tiered tagging, conditional random fields (CRF) and features tailored for inflective languages written in WCCL. The algorithm and code are inspired by Wrocław Memory-Based Tagger (WMBT) . **Availability:** download **CLARIN Centre:** CLARIN-PL **Related publication:** Radziszewski (2013)
WMBT (Wrocław Memory-Based Tagger) **Functionality:** MSD **Licence:** GNU LGPL 3.0 (info missing in the CLARIN-PL entry)	Polish	This tool uses the TiMBL API as the underlying memory-based learning implementation. The features for classification are generated by using the WCCL formalism. The tool uses a tiered tagging approach. Grammatical class is disambiguated first, then subsequent attributes (as defined in a config file) are taken care of. Each attribute may be supplied a different set of features. The software package comes with default configurations for KIPI/IPIC and NKJP tagsets. **Availability:** download **Input:** various, default is XCES XML **Output:** various, default is XCES XML **CLARIN Centre:** CLARIN-PL
Lemmatizer for Portuguese **Functionality:** lemma **Licence:** Apache Licence 2.0 (academic)	Portuguese	This tool is based on the MXPOST part of speech tagger and is trained on UNITEX dictionaries for Portuguese. **Availability:** download **Input:** plain text **Output:** plain text **CLARIN Centre:** PORTULAN
LX-Tagger **Functionality:** MSD **Licence:** Academic - Non-Commercial use	Portuguese	This tool is based on the TnT tagger (Brants 2000). **Availability:** download **CLARIN Centre:** PORTULAN

		Related publication: Silva (2007)
LX-Verbal Lemmatizer **Functionality:** lemma (verbs) **Licence:** Terms of Service	Portuguese	This tool performs fully-fledged lemmatization of Portuguese verbs, including the full range of pronominal conjugation forms. **Availability:** web application **CLARIN Centre:** PORTULAN
OpenNLP Part-of-Speech Tagger (Portuguese) **Functionality:** PoS **Licence:** Apache Licence 2.0 (restricted)	Portuguese	This tool is based on the Apache OpenNLP library, which is a perception and maximum entropy–based machine learning toolkit for the processing of natural language text. **Availability:** web application **Input:** application/xml **Output:** application/xml **CLARIN Centre:** CLARIN:EL
Character-level part-of-speech tagger of Slovene language **Functionality:** PoS **Licence:** GNU General Public Licence, version 3	Slovenian	This tool uses convolutional and LSTM neural networks. The tool has been trained on the ssj500k 2.1 corpus. **Availability:** download **Input:** XML, TEI, plain text **CLARIN Centre:** CLARIN.SI **Related publication:** Belej (2018)
janes-tagger **Functionality:** PoS, lemma	Slovenian	This tool, which was developed in the context of the JANES project, tags non-standard Slovenian, with Croatian and Serbian to follow. **Availability:** download **Input:** plain text **CLARIN Centre:** CLARIN.SI

1.1. Tools for tagging/lemmatizing within multiple languages

Table 2: PoS/MSD-taggers and lemmatizers for multiple languages, sorted by language and corpus name

Tool	Languages	Description
NCHLT Tagger **Functionality:** PoS, phrase chunks, NE **Licence:** CC-BY 2.5 South Africa Licence	Afrikaans, English, Ndebele, Xhosa, Zulu, Sesotho sa Leboa, Setswana, Sesotho, Siswati, Tshivenda, Xitsonga	This tool is used to annotate texts in Afrikaans and a variety of Bantu languages. **Availability:** download **CLARIN Centre:** SADiLaR

CST's lemmatizer **Functionality:** lemma	Bulgarian, Czech, Danish, Dutch, English, Estonian, Farsi, French, German, Greek, Hungarian, Icelandic, Italian, Latin, Macedonian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovene, Spanish, Ukrainian	This tool uses affix rules (affix: prefix, infix, suffix, circumfix). **Availability:** download **CLARIN Centre:** LINDAT/CLARIN-DK **Related publication:** Jongejan and Dalianis (2009)
Sparv **Functionality:** PoS, MSD, lemma, compound analysis, dictionary lookup	Bulgarian, English, Estonian, Finnish, French, Galician, Italian, Catalan, Latin, Dutch, Norwegian, Polish, Portuguese, Romanian, Russian, Slovak, Slovenian, Spanish, Swedish, German	This tool is Språkbanken's corpus annotation pipeline infrastructure. The pipeline uses in-house and external tools on the text to segment it into sentences and paragraphs, tokenise, tag parts-of-speech, look up in dictionaries and analyse compounds. The pipeline can also be run using a web API with XML results, and it is run locally to prepare the documents in Korp, which is SWE-LANG's corpus search tool. While the most sophisticated support is for modern Swedish, the pipeline supports additional 19 languages. **Availability:** web application, web API **Input:** plain text, XML **Output:** plain text, XML **CLARIN Centre:** SWE-CLARIN **Related publication:** Borin et al. (2016)
ReLDIanno **Functionality:** PoS, lemma, NER, syntactic parsing **Licence:** CC-BY (for webservice); Apache 2 for library	Croatian, Serbian, Slovenian	This tool, which was developed in the context of the ReLDI project, employs the MULTEXT tagset for part of speech tagging and Universal Dependencies for syntactic parsing. **Availability:** download, web application **Input:** plain text, TCF **Output:** vertical, plain text **CLARIN Centre:** CLARIN.SI

		Related publication: Ljubešić et al. (2016)
CLARIN DK NLP Toolbox **Functionality:** PoS, lemma, frequency lists	Danish, English	This tool is an NLP toolchain that is part of the core CLARIN-DK structure. **Availability:** web application **Input:** plain text, rtf, pdf **Output:** plain text, rtf **CLARIN Centre:** CLARIN-DK
GENIA Tagger **Functionality:** PoS, lemma, chunks, named entities **Licence:** proprietary - commercial	English, Czech, Slovak	This tool is used for annotating biomedical texts such as MEDLINE abstracts. **Availability:** download **CLARIN Centre:** PORTULAN **Related publication:** Tsurouka et al. (2015)
MorphoDiTa: Morphological Dictionary and Tagger **Functionality:** MSD, lemma **Licence:** Mozilla Public Licence 2.0 (software); CC BY-NC-SA models	English, Czech, Slovak	This tool performs morphological analysis, morphological generation, tagging and tokenization and is distributed as a standalone tool or a library, along with trained linguistic models. For Czech, the tool achieves state-of-the-art results with a throughput around 10-200,000 words per second. The tool is versioned using Semantic Versioning. The following language models are available through LINDAT under the CC BY licence: Czech and English. **Availability:** download, web application, API **Input:** plain text, vertical **Output:** vertical, XML **CLARIN Centre:** LINDAT **Related publication:** Straková, Straka and Hajič (2014)
STEPP Tagger **Functionality:** PoS **Licence:** proprietary - commercial	English, Czech, Slovak	This tool is used for annotating biomedical texts such as MEDLINE abstracts. **Input:** plain text **Output:** plain text **CLARIN Centre:** PORTULAN
Stanford Phrase Structure Parser **Functionality:** PoS, syntactic parsing	English, German	This tool is a Weblicht implementation of the Stanford Parser. **Availability:** WebLicht **Input:** plain text, pdf, rtf, XML **Output:** plain text, pdf, rtf, XML **CLARIN Centre:** CLARIN-D

		Related publication: Hinrichs et al. (2010)
RFTagger **Functionality:** PoS	German, Czech, Slovene, Hungarian	This tool is a PoS tagger implemented in WebLicht. **Availability:** download, WebLicht **CLARIN Centre:** CLARIN-D **Related publication:** Schmid and Laws (1995)
Sticker part-of-speech tagger UD **Functionality:** PoS, syntactic parsing, NER **Licence:** Blue Oak Mode Licence version 1.0.0	German, Dutch	This tool is a PoS tagger, syntactic parser and named entity recognizer implemented in WebLicht. The PoS tagger uses the Universal Dependencies tagset. **Availability:** download, WebLicht **CLARIN Centre:** CLARIN-D **Related publication:** Ling et al. (2015)
TreeTagger **Functionality:** PoS, lemma **Licence:** free but unspecified	German, English French, Italian, Dutch, Spanish, Bulgarian, Russian, Greek, Portuguese, Chinese, Swahili, Latin, Estonian and old French	This tool is a PoS tagger and lemmatizer implemented in WebLicht. **Availability:** download, WebLicht **Output:** plain text **CLARIN Centre:** CLARIN-D **Related publication:** Schmid (19992)
PoS Tagger OpenNLP Project **Functionality:** PoS	German, English, Italian	This tool is a PoS tagger implemented in WebLicht. The model for Italian is trained on the MIDT corpus. **Availability:** WebLicht **Input:** TCF, XML **CLARIN Centre:** CLARIN-D
UDPipe **Functionality:** PoS, lemma, syntactic parsing **Licence:** Mozilla Public Licence 2.0 (software); CC BY-NC-SA UD models	Language independent	This tool is a trainable pipeline for annotating CoNLL-U files. UDPipe is language-agnostic and can be trained given annotated data in the CoNLL-U format. Trained models are provided for nearly all Universal Dependency treebanks. **Availability:** download, web application **Input:** plain text **Output:** CoNLL-U **CLARIN Centre:** LINDAT **Related publication:** Straka and Straková (2017)

Turku-neural-parser-pipeline **Functionality:** segmentation, MSD, syntactic parsing, lemma **Licence:** Apache License 2.0	More than 50 languages	A neural parsing pipeline for segmentation, morphological tagging, dependency parsing and lemmatization with pre-trained models for more than 50 languages. Top ranker in the CoNLL-18 Shared Task. **Availability:** download, web application **Input:** utf-8 encoded plain text **Output:** CoNLL-U **CLARIN Centre:** FIN-CLARIN **Related publication:** Kanerva et al. (2018)

3. The overview

3.1. Identification

54 (83%) out of 65 taggers have VLO entries, except for the following 11 (in the parentheses, we specify the CLARIN consortium relevant to the tool):

(1) ReLDIanno (CLARIN.SI)
(2) janes-tagger (CLARIN.SI)
(3) CLARIN DK NLP Toolbox (CLARIN-DK)
(4) ILSP Feature-based multi-tiered POS Tagger (CLARIN:EL)
(5) OpenNLP Part-of-Speech Tagger (English) (CLARIN:EL)
(6) OpenNLP Part-of-Speech Tagger (German) (CLARIN:EL)
(7) OpenNLP Part-of-Speech Tagger (Portuguese) (CLARIN:EL)
(8) Sparv (SWE-CLARIN)
(9) CLaRK (CLaDA-BG)
(10) NLP-PIPE (CLARIN-LV)
(11) Turku-neural-parser-pipeline (FIN-CLARIN)

Curation is needed for the VLO entries of the following 6 tools (all of which are part of the *WebLicht Webservice Orchestrator* collection in the VLO) that are otherwise accessible through the WebLicht environment:

(1) Stanford Dependency Parser
(2) Stanford Phrase Structure Parser
(3) Stuttgart Dependency Parser
(4) WebLicht Part-of-Speech Tagger
(5) POS Tagger OpenNLP Project
(6) MorphAdorner Lemmatizer

The handle links listed in the VLO for each tool resolve into a metadata CMDI XML file which isn't user friendly. However, since the relevant link to the WebLicht application or any other relevant access point is not included in the entries, none of these tools can actually be directly accessed through the VLO.

3.2. Availability

In this section we list the availability of the tools and highlight those that seem to be unavailable.

As shown below, 11 (17%) tools are available both for download and as web applications, 29 (44%) tools are available only for download, 18 (28%) tools are available as web applications, and 7 (11%) tools are unavailable. In the parentheses, we list the CLARIN consortium/observer which provides/lists the tool.

3.2.1. Download and web application

(1) ReLDIanno (CLARIN.SI)
(2) MorphoDiTa: Morphological Dictionary and Tagger (LINDAT)
(3) UDPipe (LINDAT)
(4) FinTag (FIN-CLARIN)
(5) Morfeusz 2 (CLARIN-PL)
(6) RFTagger (CLARIN-D)
(7) TreeTagger (CLARIN-D)
(8) Sticker part-of-speech tagger UD (CLARIN-D)
(9) Vabamorf open source morphology tagger for Estonian (CELR)
(10) IceNLP Natural Language Processing toolkit (CLARIN-IS)
(11) Turku-neural-parser-pipeline (FIN-CLARIN)

3.2.2. Download

(1) Character-level part-of-speech tagger of Slovene language (CLARIN.SI)
(2) janes-tagger (CLARIN.SI)
(3) LX-Tagger (PORTULAN)
(4) GENIA Tagger (PORTULAN)
(5) NCHLT Tagger (SADiLaR)
(6) HMM tagger (LINDAT)
(7) Hunpos (LINDAT)
(8) The Oslo-Bergen tagger (CLARINO)
(9) Frog (CLARIAH-NL)
(10) Tagger SentiOne - version 2 (CLARIN-PL)
(11) MorphoDiTa-based tagger for Polish language (CLARIN-PL)
(12) WCRFT (CLARIN-PL)
(13) WMBT (CLARIN-PL)
(14) CLaRK (CLaDA-BG)
(15) EstNLTK (CELR)
(16) NLP-PIPE (CLARIN-LV)
(17) NCHLT Sepedi Lemmatiser (SADiLaR)
(18) NCHLT Sesotho Lemmatiser (SADiLaR)

(19) NCHLT Setswana Lemmatiser (SADiLaR)
(20) NCHLT Siswati Lemmatiser (SADiLaR)
(21) NCHLT isiZulu Lemmatiser (SADiLaR)
(22) NCHLT isiXhosa Lemmatiser (SADiLaR)
(23) NCHLT isiNdebele Lemmatiser (SADiLaR)
(24) NCHLT Afrikaans Lemmatiser (SADiLaR)
(25) NCHLT Tshivenda Lemmatiser (SADiLaR)
(26) NCHLT Xitsonga Lemmatiser (SADiLaR)
(27) Lemmatizer for Portuguese (PORTULAN)
(28) CST's lemmatizer (LINDAT/CLARIN-DK)
(29) TaKIPI (CLARIN-PL)

3.2.3. Web application

(1) Tagger WS (CLARIN-PL)
(2) CLARIN DK NLP Toolbox (CLARIN-DK)
(3) ILSP Feature-based multi-tiered POS Tagger (CLARIN:EL)
(4) OpenNLP Part-of-Speech Tagger (English) (CLARIN:EL)
(5) OpenNLP Part-of-Speech Tagger (German) (CLARIN:EL)
(6) OpenNLP Part-of-Speech Tagger (Portuguese) (CLARIN:EL)
(7) Stanford Dependency Parser (CLARIN-D, WebLicht)
(8) Stanford Phrase Structure Parser (CLARIN-D, WebLicht)
(9) Stuttgart Dependency Parser (CLARIN-D, WebLicht)
(10) Part-of-Speech Tagger (CLARIN-D, WebLicht)
(11) POS Tagger OpenNLP Project (CLARIN-D, WebLicht)
(12) LX-Verbal Lemmatizer (PORTULAN)
(13) ILSP lemmatizer (CLARIN:EL)
(14) INL Labs tagger/lemmatizer tools (CLARIAH-NL)
(15) Freeling (CLARIN-IT)
(16) SepVerb Lemmatizer (CLARIN-D)
(17) MorphAdorner Lemmatizer (CLARIN-D)
(18) Corpus.by Lemmatizer (CLARIN Knowledge Centre for Belarusian text and speech processing)

3.2.4. Unavailable

(1) STEPP Tagger (PORTULAN)
(2) Afrikaans TnT-Tagger (SADiLaR)
(3) Assamese POS Tagger (CLARIN-PL)
(4) Sepedi Part of Speech Tagger (SADiLaR)
(5) MLSS Tagger Web Service (PORTULAN)
(6) Tadpole (LINDAT/CLARIAH-NL)
(7) SMOR lemmatizer (CLARIN-D)

Unavailability is due to at least three reasons. First, STEPP Tagger is likely not available for download or online browsing due to its proprietary/commercial licence. Second, the tools

Tadpole and SMOR lemmatizer are not available partially because the links to their external landing pages are broken – see http://ilk.uvt.nl/tadpole for Tadpole and http://hdl.handle.net/11022/1007-0000-0000-8E46-2 for SMOR lemmatizer (last checked 27 March 2020). Third, the link to the external landing page provided for MLSS Tagger Web Service resolves to a metadata description rather than a website where the tool can be accessed, while the tool cannot be downloaded from the PORTULAN repository. By contrast, it is unclear why the remaining tools – i.e., Afrikaans TnT-Tagger, Assamese POS Tagger, Sepedi Part of Speech Tagger – are unavailable.

3.3. Metadata

3.3.1. Language

50 (77%) tools are aimed at processing a single language:

(1) Afrikaans (2 tool)
(2) Assamese (1 tool)
(3) Bantu languages (10 tools)
(4) Belarussian (1 tool)
(5) Bulgarian (1 tool)
(6) Czech (1 tool)
(7) Dutch (3 tools)
(8) English (3 tools)
(9) Estonian (2 tools)
(10) Finnish (1 tool)
(11) German (5 tools)
(12) Greek (1 tools)
(13) Hungarian (1 tool)
(14) Icelandic (1 tool)
(15) Italian (1 tool)
(16) Latvian (1 tool)
(17) Maltese (1 tool)
(18) Norwegian (1 tool)
(19) Polish (7 tools)
(20) Portuguese (4 tools)
(21) Slovenian (2 tools)

15 (23%) tools are aimed at processing multiple languages. Especially noteworthy in this respect are the tools UDPipe and Turku-neural-parser-pipeline, both of which make available pre-trained models for more than 50 languages. See Universal Dependencies 2.5 Models for UDPipe and the 82 CoNLL-18 Shared Task models for the Turku parser.

3.3.2. Functionality

The following 22 (34%) tools perform part-of-speech tagging or morphosyntactic annotation only:

(1) Character-level part-of-speech tagger of Slovene language
(2) LX-Tagger
(3) STEPP Tagger
(4) HMM tagger
(5) Afrikaans TnT-Tagger
(6) hunpos
(7) Assamese POS Tagger
(8) Tagger SentiOne - version 2
(9) MorphoDiTa-based tagger for Polish language
(10) WCRFT
(11) WMBT
(12) TaKIPI
(13) ILSP Feature-based multi-tiered POS Tagger
(14) OpenNLP Part-of-Speech Tagger (English)
(15) OpenNLP Part-of-Speech Tagger (German)
(16) OpenNLP Part-of-Speech Tagger (Portuguese)
(17) Morfeusz 2
(18) RFTagger
(19) WebLicht Part-of-Speech Tagger
(20) POS Tagger OpenNLP Project
(21) Sepedi Part of Speech Tagger
(22) MLSS Tagger Web Service

The following 17 (26%) tools perform lemmatization only:

(1) LX-Verbal Lemmatizer
(2) ILSP lemmatizer
(3) NCHLT Sepedi Lemmatiser
(4) NCHLT Sesotho Lemmatiser
(5) NCHLT Setswana Lemmatiser
(6) NCHLT Siswati Lemmatiser
(7) NCHLT isiZulu Lemmatiser
(8) NCHLT isiXhosa Lemmatiser
(9) NCHLT isiNdebele Lemmatiser
(10) NCHLT Afrikaans Lemmatiser
(11) NCHLT Tshivenda Lemmatiser
(12) NCHLT Xitsonga Lemmatiser
(13) Lemmatizer for Portuguese
(14) CST's lemmatizer
(15) SepVerb Lemmatizer

(16) MorphAdorner Lemmatizer
(17) Corpus.by Lemmatizer

The following 26 (40%) tools perform multiple tasks which are listed in the parentheses:

(1) ReLDIanno (*PoS/MSD-tagging*, *lemmatization*, *named entity recognition*, *syntactic parsing*)
(2) MorphoDiTa: Morphological Dictionary and Tagger (*PoS/MSD-tagging*, *lemmatization*)
(3) UDPipe (*PoS/MSD-tagging*, *lemmatization*, *syntactic parsing*)
(4) GENIA Tagger (*PoS/MSD-tagging*, *lemmatization*, *phrase chunking*, *named entity recognition*)
(5) NCHLT Tagger (*PoS/MSD-tagging*, *phrase chunking, named entity recognition*)
(6) Tagger WS (*PoS/MSD-tagging*, *lemmatization*)
(7) The Oslo-Bergen tagger (*PoS/MSD-tagging*, *syntactic parsing*)
(8) Frog (*PoS/MSD-tagging*, *lemmatization*, *named entity recognition*, *phrase chunking*, *syntactic parsing/dependency relations with headwords*)
(9) FinTag (*PoS/MSD-tagging*, *lemmatization*, *named entity recognition*)
(10) CLARIN DK NLP Toolbox (*PoS/MSD-tagging*, *lemmatization*, *frequency lists*)
(11) Sparv (*PoS/MSD-tagging*, *lemmatization*)
(12) CLaRK (*PoS/MSD-tagging*, *lemmatization*, *syntactic parsing*)
(13) Stanford Dependency Parser (*PoS/MSD-tagging*, *syntactic parsing*)
(14) Stanford Phrase Structure Parser (*PoS/MSD-tagging*, *syntactic parsing*)
(15) Stuttgart Dependency Parser (*PoS/MSD-tagging*, *lemmatization*, *syntactic parsing*)
(16) TreeTagger (*PoS/MSD-tagging*, *lemmatization*)
(17) Sticker part-of-speech tagger UD (*PoS/MSD-tagging*, *syntactic parsing*, *named entity recognition*)
(18) EstNLTK (*PoS/MSD-tagging*, *named entity recognition*)
(19) Vabamorf open source morphology tagger for Estonian (*PoS/MSD-tagging*, *lemmatization*)
(20) NLP-PIPE (*PoS/MSD-tagging*, *syntactic parsing*, *named entity recognition*)
(21) IceNLP Natural Language Processing toolkit (*PoS/MSD-tagging*, *lemmatization*, *syntactic parsing*)
(22) Turku-neural-parser-pipeline (*PoS/MSD-tagging*, *lemmatization*, *syntactic parsing*)
(23) SMOR lemmatizer (*PoS/MSD-tagging*, *lemmatization*)
(24) Freeling (*PoS/MSD-tagging*, *lemmatization*)
(25) INL Labs tagger/lemmatizer tools (*PoS/MSD-tagging*, *lemmatization*)
(26) Tadpole (*PoS/MSD-tagging*, *lemmatization*, *syntactic parsing*)

In sum, these are the most common tasks performed by the overviewed tools:

(1) PoS/MSD-tagging (48 tools, 74%)
(2) Lemmatization (38 tools, 58%)
(3) Syntactic parsing (17 tools, 26%)

3.3.3. Input/output format

More than half (41, 63%) of the tools do not provide information on the input and output formats of the text or do so only partially.

(1) NCHLT Sepedi Lemmatiser
(2) NCHLT Sesotho Lemmatiser
(3) NCHLT Setswana Lemmatiser
(4) NCHLT Siswati Lemmatiser
(5) NCHLT isiZulu Lemmatiser
(6) NCHLT isiXhosa Lemmatiser
(7) NCHLT isiNdebele Lemmatiser
(8) NCHLT Afrikaans Lemmatiser
(9) NCHLT Tshivenda Lemmatiser
(10) NCHLT Xitsonga Lemmatiser
(11) Tadpole
(12) Freeling
(13) LX-Tagger
(14) HMM tagger
(15) hunpos
(16) The Oslo-Bergen tagger
(17) Assamese POS Tagger
(18) Tagger SentiOne - version 2
(19) MorphoDiTa-based tagger for Polish language
(20) WCRFT (Wrocław CRF Tagger)
(21) TaKIPI
(22) Sepedi Part of Speech Tagger
(23) MLSS Tagger Web Service
(24) NLP-PIPE
(25) LX-Verbal Lemmatizer
(26) CST's lemmatizer
(27) RFTagger
(28) GENIA Tagger
(29) STEPP Tagger
(30) NCHLT Tagger
(31) INL Labs tagger/lemmatizer tools
(32) SMOR lemmatizer
(33) SepVerb Lemmatizer
(34) MorphAdorner Lemmatizer
(35) Character-level part-of-speech tagger of Slovene language
(36) janes-tagger
(37) Frog
(38) WebLicht Part-of-Speech Tagger
(39) Sticker part-of-speech tagger UD
(40) PoS Tagger OpenNLP Project

(41) TreeTagger

The following 24 (37%) tools specify the types of the input and output files:

(1) Lemmatizer for Portuguese (**input:** plain text, **output:** plain text)
(2) Corpus.by Lemmatizer (**input:** plain text, **output:** plain text)
(3) Tagger WS (**input:** plain text, **output:** XML)
(4) Afrikaans TnT-Tagger (**input:** plain text, **output:** plain text)
(5) WMBT (Wrocław Memory-Based Tagger) (**input:** XCES XML, various, **output:** XCES XML, various)
(6) FinTag (**input:** plain text, pdf, doc, scv, epub, html, odt, xls, **output:** TSV)
(7) ILSP Feature-based multi-tiered POS Tagger (**input:** XMI, XML, **output:** XMI, XML)
(8) OpenNLP Part-of-Speech Tagger (English) (**input:** XML, **output:** XML)
(9) OpenNLP Part-of-Speech Tagger (German) (**input:** XML, **output:** XML)
(10) OpenNLP Part-of-Speech Tagger (Portuguese) (**input:** XML, **output:** XML)
(11) Morfeusz 2 (**input:** plain text, various, **output:** plain text, various)
(12) CLaRK (**input:** XML, **output:** XML)
(13) Stanford Dependency Parser (**input:** plain text, pdf, rtf, XML, **output:** plain text, pdf, rtf, XML)
(14) Stuttgart Dependency Parser (**input:** plain text, pdf, rtf, XML, **output:** plain text, pdf, rtf, XML)
(15) EstNLTK (**input:** plain text, **output:** plain text)
(16) Vabamorf open source morphology tagger for Estonian (**input:** plain text, **output:** plain text)
(17) IceNLP Natural Language Processing toolkit (**input:** plain text, **output:** plain text)
(18) Turku-neural-parser-pipeline (**input:** utf-8 encoded plain text, **output:** CoNLL-U)
(19) Sparv (**input:** plain text, XML, **output:** plain text, XML)
(20) ReLDIanno (**input:** Plain text, TCF, **output:** vertical, plain text)
(21) UDPipe (**input:** plain text, **output:** CoNLL-U)
(22) MorphoDiTa: Morphological Dictionary and Tagger (**input:** plain text, vertical, **output:** vertical, XML)
(23) CLARIN DK NLP Toolbox (**input:** plain text, rtf, pdf, **output:** plain text)
(24) Stanford Phrase Structure Parser (**input:** plain text, pdf, rtf, XML, **output:** plain text, pdf, rtf, XML)

3.3.4. Licence

Information on licence is included for 43 tools (66%) and is missing for the following 22 tools:

(1) janes-tagger (CLARIN.SI)
(2) Tagger WS (CLARIN-PL)
(3) Afrikaans TnT-Tagger (SADiLaR)
(4) Assamese POS Tagger (CLARIN-PL)
(5) TaKIPI (CLARIN-PL)
(6) CLARIN DK NLP Toolbox (CLARIN-DK)

(7) Sparv (SWE-CLARIN)
(8) CLaRK (CLaDA-BG)
(9) Stanford Dependency Parser (CLARIN-D)
(10) Stanford Phrase Structure Parser (CLARIN-D)
(11) Stuttgart Dependency Parser (CLARIN-D)
(12) Part-of-Speech Tagger (CLARIN-D)
(13) POS Tagger OpenNLP Project (CLARIN-D)
(14) Sepedi Part of Speech Tagger (SADiLaR)
(15) LX-Verbal Lemmatizer (PORTULAN)
(16) Tadpole (LINDAT/CLARIAH-NL)
(17) Freeling (CLARIN-IT)
(18) SMOR lemmatizer (CLARIN-D)
(19) CST's lemmatizer (CLARIN-DK)
(20) SepVerb Lemmatizer (CLARIN-D)
(21) MorphAdorner Lemmatizer (CLARIN-D)
(22) Corpus.by Lemmatizer (CLARIN Knowledge Centre for Belarusian text and speech processing)

The most common licences are as follows:

(1) CC-BY (16 tools)
(2) GNU General Public Licence (10 tools)
(3) Apache restricted/academic (6 tools)

It is worth noting that in the case of the tools WCRFT (Wrocław CRF Tagger) and WMBT (Wrocław Memory-Based Tagger), the relevant licence (i.e., GNU LGPL 3.0 for both tools) is only specified on the external landing page, but is missing from the entry in the CLARIN-PL repository.

4. Conclusion

We have provided an overview of 65 tools for part of speech tagging or lemmatization and evaluated them from the perspective of VLO findability and availability, as well as from the perspective of the metadata describing language, functionality, and licence.

In relation to findability, 54 of the 65 (83%) tools are available in the VLO, which is better than the tools for text normalization (29%) and tools for named entity recognition (79%). The VLO entries of 6 tools that are available through WebLicht are suboptimally presented, since they link only to metadata descriptions that aren't user friendly. It seems that to varying degrees this is an issue of most VLO entries that are part of the *WebLicht Webservice Orchestrator* collection and would warrant manual curation (see also p. 9 of the report on tools of named entity recognition).

In relation to accessibility, 11 (17%) tools are available both for download and as web applications, 29 (44%) tools are available only for download, 18 (28%) tools are available as

web applications, and 8 (11%) tools are unavailable. Crucially, unavailability in two cases is due to broken links to external landing pages, which warrants curation.

In relation to language, the majority of the (50/77%) tools are aimed at pre-processing within a single language. Especially noteworthy is the fact that 10 tools are aimed at lemmatizing within Bantu languages spoken in South Africa, like Sesotho and Zulu.

In terms of functionality, 22 (34%) tools are dedicated exclusively to PoS/MSD-tagging, 17 (26%) tools exclusively perform lemmatization, while 26 (40%) tools perform several tasks apart from PoS/MSD-tagging or lemmatization, such as named entity recognition and syntactic parsing.

Finally, licence is included for 43 tools (66%).

5. References

Barnard, Etienne, Marelie H. Davel, Charl van Heerden, Febe de Wet, and Jaco Badenhors. 2014. The NCHLT Speech Corpus of the South African languages. In *SLTU-2014*, 194–200. http://mica.edu.vn/sltu2014/proceedings/28.pdf.

Belej, Primož. 2018. *Oblikoskladenjsko označevanje slovenskega jezika z globokimi nevronskimi mrežami*. Master's Thesis. University of Ljubljana. https://repozitorij.uni-lj.si/IzpisGradiva.php?id=105266&lang=eng.

Borin, Lars, Markus Forsberg, Martin Hammarstedt, Dan Rosén, Roland Schäfer, Anne Schumacher. 2016. Sparv: Språkbanken's corpus annotation pipeline infrastructure. In *Proceedings of SLTC 2016*. https://gup.ub.gu.se/publication/246053?lang=sv.

Bosch, Antal van den, Bertjan Busser, Sander Canisius, and Walter Daelemans. 2007. An efficient memory-based morphosyntactic tagger and parser for Dutch, In *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting*, edited by F. van Eynde, P. Dirix, I. Schuurman, and V. Vandeghinste, 99–114. http://hdl.handle.net/1874/296756.

Brants, Thorsten. 2000. TnT – A Statistical Part of-Speech Tagger. http://www.coli.uni-saarland.de/~thorsten/publications/Brants-TR-TnT.pdf.

Halácsy, Péter, Andras Kornai, and Csaba Oravecz. 2007. HunPos: an open source trigram tagger.https://www.researchgate.net/publication/228524009_HunPos_an_open_source_trigram_tagger.

Hinrichs, Erhard, Marie Hinrichs, and Thomas Zastrow. 2010. WebLicht: Web-Based LRT Services for German. In *Proceedings of the ACL 2010 System Demonstrations*, 25–29. https://www.aclweb.org/anthology/P10-4005.

Johannessen, Janne Bondi, Kristin Hagen, André Lynum, and Anders Nøklestad. 2012. A combined rule-based and statistical tagger. In *Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian*, edited by G. Andersen, 51–66. https://doi.org/10.1075/scl.49.03joh.

Jongejan, Bart, and Hercules Dalianis. 2009. Automatic training of lemmatization rules that handle morphological changes in pre-, in-and suffixes alike. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, 145–153. https://www.aclweb.org/anthology/P09-1017.pdf.

Kaalep, Heiki-Jaan. 2015. Vabamorf, a set of open-source morphological tools for Estonian. https://dh.org.ee/events/dh2015/abstracts/kaalep/.

Kanerva, Jenna, Filip Ginter, Niko Miekka, Akseli Leino, and Tapio Salakosk. 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*, 133–142. https://www.aclweb.org/anthology/K18-2013.pdf.

Ling, Wang, Chris Dyer, Alan W Black, Isabel Trancoso, Ramón Fermandez, Silvio Amir, Luís Marujo, and Tiago Luís. 2015. Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1520–1530. http://dx.doi.org/10.18653/v1/D15-1176.

Ljubešić, Nikola, Filip Klubička, Željko Agić, and Ivo-Pavao Jazbec. 2016. New Inflectional Lexicons and Training Corpora for Improved Morphosyntactic Annotation of Croatian and Serbian. In *Proceedings of LREC 2016*, edited by Nicoletta Calzolari, 4264–4270. http://www.lrec-conf.org/proceedings/lrec2016/pdf/340_Paper.pdf.

Loftsson, Hrafn, and Eiríkur Rögnvaldsson. 2007. IceNLP: A natural language processing toolkit for Icelandic. In *Proceedings of the Eighth Annual Conference of the International Speech Communication Association*. http://www.ru.is/kennarar/hrafn/papers/IceNLP_final2.pdf.

Orasmaa, Siim, Timo Petmanson, Alexander Tkachenko, Sven Laur, and Heiki-Jaan Kaalep. 2016. Estnltk-nlp toolkit for Estonian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2460–2466. https://www.aclweb.org/anthology/L16-1390.pdf.

Padró, Lluís, Miquel Colaldo, Samuel Reese, Marina Lloberes, and Irene Castellón. 2010. FreeLing 2.1. Five Years of open-source language processing tools. In *Proceedings of LREC2010*, 931–936. http://www.lrec-conf.org/proceedings/lrec2010/pdf/14_Paper.pdf.

Papageorgiou, Harris, Prokopis Prokopidis, Voula Giouli, and Stelios Piperidis. 2000. A Unified POS Tagging Architecture and its Application to Greek. In *Proceedings of LREC2000*. http://www.lrec-conf.org/proceedings/lrec2000/pdf/181.pdf.

Piasecki, Maciej. 2007. Polish tagger TaKIPI: Rule based construction and optimisation. *Task quarterly*, 11 (1–2): 151–167. https://153.19.250.1/files/quart/TQ2007/01-02/tq111t-g.pdf.

Prokopidis, Prokopis, Byron Georgantopoulos, and Haris Papageorgiou. 2011. A Suite of Natural Language Processing Tools for Greek. In *The 10th International Conference of Greek Linguistics*. http://nlp.ilsp.gr/nlp/ICGL2011_Prokopidis_etal.pdf.

Przepiórkowski, Adam. 2005. The IPI PAN Corpus in numbers. In *Proceedings of the 2nd Language & Technology Conference*, 27–31. https://www.researchgate.net/publication/228377475_The_IPI_PAN_Corpus_in_numbers.

Radziszewski, Adam. 2013. A Tiered CRF Tagger for Polish. In *Intelligent Tools for Building a Scientific Information Platform*, 215–230. https://doi.org/10.1007/978-3-642-35647-6_16.

Schmid, Helmut, and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, 777–784. https://www.aclweb.org/anthology/C08-1098.pdf.

Schmid, Helmut. 1994. Probabilistic part-ofispeech tagging using decision trees. In *New methods in language processing*. https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf.

Schmid, Helmut. 1999. Improvements in part-of-speech tagging with an application to German. In *Natural language processing using very large corpora*, 13–25. Springer, Dordrecht. https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger2.pdf.

Silva, João. 2007. *Shallow Processing of Portuguese: From Sentence Chunking to Nominal Lemmatization*. Master's Thesis.

Simov, Kiril, Zdravko Peev, Milen Kouylekov, Alexander Simov, Marin Dimitrov, and Atanas Kiryakov. 2017. ClaRK – an XML-based System for Corpora Development. In *Proc. of the Corpus Linguistics 2001 Conference*, 558–560. http://bultreebank.org/wp-content/uploads/2017/04/BTB-TR06.pdf.

Straka, Milan, and Jana Straková. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. http://ufal.mff.cuni.cz/~straka/papers/2017-conll_udpipe.pdf.

Straková, Jana, Milan Straka, and Jan Hajič. 2014. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 13–18. https://www.aclweb.org/anthology/P14-5003.pdf.

Tsuruoka, Yoshimasa, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Jun'ichi Tsujii. 2005. Developing a Robust Part-of-Speech Tagger for Biomedical Text. In *Advances in Informatics. PCI 2005. Lecture Notes in Computer Science*, edited by P. Bozanis and E.N. Houstis. https://doi.org/10.1007/11573036_36.

Woliński, Marcin. 2014. Morfeusz reloaded. In *Proceedings of LREC2014*, 1106–1111. http://nlp.ipipan.waw.pl/Bib/wol:14.pdf.

Znotiņš, Artūrs, and Elita Cīrule. 2018. NLP-PIPE: Latvian NLP Tool Pipeline. In *Human Language Technologies – The Baltic Perspective*, 183–189. https://doi.org/10.3233/978-1-61499-912-6-183.