

The Durm German Lemmatizer

- [Durm](#)
- [Durm German Lemmatizer](#)
- [GATE Components](#)
- [German](#)
- [Lemmatization](#)

toc_collapse=0; Table of Contents

- [1. Overview](#)
- [2. Prerequisites](#)
- [3. Documentation](#)
 - [3.1. Quick Start Guide](#)
 - [3.2. The Lexicon](#)
- [4. Evaluation](#)
- [5. Download](#)
- [6. Version history](#)
- [7. Feedback](#)
- [8. Acknowledgments](#)

The screenshot shows the Durm German Lemmatizer interface in the 'Text' tab. The main area displays a block of German text with various words highlighted in blue, indicating they have been annotated. A toolbar at the top includes buttons for 'Annotation Sets', 'Annotations', 'Co-reference Editor', and 'Text'. Below the text, a table lists 1674 annotations, each with a 'Type', 'Set', 'Start', and 'End' column, and a detailed description in the last column. To the right, a sidebar lists various annotation types with corresponding color-coded boxes: DATE (red), DE-Morph (red), DEFAULT_TOKEN (yellow), Exception (red), LOC (green), Lemma (purple, checked), Lookup (purple), NMB (yellow), NP (cyan), ORG (yellow), PER (purple), PRC (green), PROF (pink), Sentence (purple), SpaceToken (red), Split (green), TIME (cyan), Token (red), tempNP (green), and a 'Manual' section. A 'New' button is located at the bottom right of the sidebar.

Type	Set	Start	End	Annotation Details
Lemma	17183	17196		{Lemma=Bundeskanzler}
Lemma	17206	17215		{Lemma=Bundestag}
Lemma	17228	17232		{Lemma=Wahl}
Lemma	17279	17288		{Lemma=Vorschlag}
Lemma	17293	17310		{Lemma=Bundespräsident}
Lemma	17439	17444		{Lemma=Abend}

1674 Annotations (0 selected)

Document Editor Initialisation Parameters

1. Overview

The *Durm German Lemmatization System* consists of a number of [GATE](#) components and resources that perform morphological analysis and lemmatization for German nouns. The approach is described in our paper: [Perera, P., and R. Witte, "A Self-Learning Context-Aware Lemmatizer for German"](#), *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, Vancouver, British Columbia, Canada : Association for Computational Linguistics (ACL), pp. 636–643, October 6–8, 2005.

It includes:

- The **Case Tagger**, which adds case information (*Nominativ, Genitiv, Dativ, Akkusativ*) to nouns;
- The **POS-based Number Tagger**, which adds number information (*singular, plural*);
- The **Morphological Analyzer**, which classifies nouns into morphological classes;
- The **Lemmatizer**, which annotates nouns with their lemma.

Additionally, it comes with two main resources:

- **Case Tagger Probabilities**, a set of resource files with statistical information for the HMM module

- **German Lexicon**, an automatically created and updated German lexicon containing lemma, number, and case information for nouns.

All components, as well as the lexicon and the evaluation corpus with manual annotations, are distributed as free/open source software under the [GNU GPL license](#).

2. Prerequisites

In order to run the components of the Durm Lemmatizer, you'll need:

- a recent version of the [GATE system](#) (3.1 or better)
- a POS-tagger for German (currently we only support the STTS tagset as used by the [TreeTagger](#)); and
- the [MuNPEx](#) noun phrase chunker for German.

In short, if you have a processing pipeline that can create proper NP (noun phrase) annotations for German documents, you're ready to go.

Note that in the following discussion we assume you know how to work with GATE, for tasks like adding a new CREOLE repository or loading new components into a processing pipeline. If you haven't done this before, please read the [GATE user's guide](#) first!

3. Documentation

3.1. Quick Start Guide

You should have a working GATE installation including the TreeTagger as outlined under [Prerequisites](#). Then:

The Durm German Lemmatizer

Published on semanticsoftware.info (<https://www.semanticsoftware.info>)

Selected Processing resources		
1	Name	Type
1	gate	Document Reader
2	tokatree	GATE Interactor Tokenizer
3	tokatree+posprocessor	Java Processor
4	german	Spont. Sentence Splitter
5	german_tagger	Tree Tagger
6	german_pos	POS Tagger
7	german_ner	ANNIE Gazetteer
8	german_npex	ANNIE Gazetteer
9	german_npex	JAPE Transducer
10	german_npex_posprocessor	JAPE Processor
11	German German-English-Phrase Extractor	JAPE Transducer
12	German Case Tagger	Case Tagger
13	German Case Tagger	Case Tagger
14	German Case Tagger	Case Tagger
15	German MorphologicalAnalyzer	GermanMorphologicalAnalyzer
16	German Lemmatizer	GermanLemmatizer
17	German Lemmatizer Token Tagger	Document Reader FG

1. Download the complete archive (source code and Java .class-files for all components), the German lexicon, and the Case Tagger probabilities file below and unpack them;
2. Load GATE's sample application for German: `german+tagger.gapp` (you can find it under `gate/plugins/german/resources/`). **Note:** this pipeline works on the annotation set "NE," so you'll either have to (a) also use "NE" as the input/output annotation set for all downstream components listed below or (b) (perhaps simpler) remove all references to "NE" within the GATE pipeline to make it work on the default annotation set.
3. Add the [MuNPEx](#) noun phrase chunker for German (using the main grammar file `de-np_main.jape`)
4. Add a "JAPE-Transducer" component with the grammar file `DeLem/de_morph_main.jape`
5. Add the *Case Tagger* component (`CaseTagger/build`). Here you'll have to set the initialization parameter to the `CaseProbs` directory containing the probability files.
Note: To add this and the next three components from the GUI, use *File ? Manage CREOLE plugins ? Add a new CREOLE repository* and then select the indicated `build` directory.
6. Add the *Number Tagger* component (`Number/build`)
7. Add the *German Morphological Analyzer* component (`GermanMorphologicalAnalyzer/build`)
8. Add the main *German Lemmatizer* component (`GermanLemmatizer/build`). Here you'll have to set the initialization parameter to the file containing the German lexicon: `DE-Lexicon/delexicon.txt`.
9. (Optionally: add a "Document Reset" component to remove the temp annotations `NPFNP`, `PosNumber`, `Preposition`, `PN`, `Case`, `Num`, and `Gender`)
10. Load some German texts and run the pipeline. Enjoy the new annotations for **Lemma** and **DE-Morph**.

More [detailed documentation](#) (pdf) is also available.

3.2. The Lexicon

The lexicon is part of the Durm Lemmatizer. It is automatically created by processing German documents. Additional functionality implemented in the *GermanLemmatizer* component allows it to self-correct some errors that are introduced by the system. Typical entries in the lexicon look like this:

Mensch	Sg	Masc	Akk	Mensch	1	4/11/2005 15:10:26	4/11/2005 15:10:26	115
unlocked								
Menschen	Sg	Masc	Akk	Mensch	6	4/11/2005 15:8:16	4/11/2005 15:10:11	116
unlocked								

The lexicon stores full forms of words with their lemma, morphological features and some additional information. Here, *Sg* means that the number of the entry is singular. *Masc* specifies that the entry's gender is masculine and *Akk* says that its case is accusative. Next, it stores the lemma. Here, you can see that the lemma of the noun *Menschen* is *Mensch*. The number *1* in the entry states that it has been found once, when generating the lexicon. The two entries with the timestamp record the insertion and last modification time of the entry. The number *115* is a reference to a document, which specifies where the entry has been found. At the end the string *unlocked* specifies that the entry's lemma can be updated automatically and has not yet been manually corrected.

4. Evaluation

We manually annotated several texts from the [German Wikipedia](#) for **Lemma**, **Number**, and **Case**. You can use them to evaluate the lemmatizer's performance ([download](#)). These texts, including our annotations, are distributed under the [GNU Free Documentation License \(FDL\)](#).

For more details on the evaluation, have a look at our paper and the GATE user's manual.

5. Download

Latest version is v1.0 from 22.02.2006. Download:

- the complete gzipped tar archive with source code and pre-build Java .class-files for all components, including the the Case Tagger parameter files, the additional JAPE grammars, and the lexicon: [DurmLemmatizer-1.0.tgz](#)
- only the [German lexicon](#) version 20060222 with 84320 entries (note: in Unicode UTF-8 encoding!)
- only the technical [documentation](#) (PDF format)
- our research paper on [German lemmatization](#)
- our [manually annotated texts](#) (distributed under the [GNU FDL](#))
- the GNU GPL [license](#), under which you can use all the components and their resources (including the lexicon)

6. Version history

- 1.0: 22.02.2006. Initial public release.

7. Feedback

For questions, comments, etc., please use the [Durm Forum](#).

8. Acknowledgments

Development of the German Lemmatizer has been supported by the German research foundation (DFG) within the project "Entstehungswissen" (LO296/18-1). The TIGER Treebank (Version 2) has been used for training and evaluation of the Case Tagger.



Except where otherwise noted, all original content on this site is copyright by its author and licensed under a [Creative Commons Attribution-Share Alike 2.5 Canada License](#).

Source URL (retrieved on 2026-01-25 23:40): <https://www.semanticsoftware.info/durm-german-lemmatizer>