

Dr. Gerd Willée
Dr. Karlheinz Stöber
IKP – Universität Bonn
Poppelsdorfer Allee 47

53115 Bonn

Germany

willee@uni-bonn.de
kst@ikp.uni-bonn.de

LEMMA3 – a Primarily Wordform Based Wordclass Tagger and Lemmatizer for Unrestricted German Texts¹

Abstract:

The functionality and structure of the word class tagger and lemmatizer LEMMA3 for unrestricted German texts is described. The program works mainly word form based with external analysis lists (full form dictionary of the elements of the closed word classes and a deflection list for nouns, verbs and adjectives); for the resolution of homographic forms context rules are used which are part of the program code. A test with a part of Thomas Mann's showed rather good results. The program is written in C++. The complete system will be freely available for scientific use.

Contents:

0. Introduction
1. Functionality
2. Structure
3. Limitations
4. Intended Use
5. Evaluation
6. Intended Use
7. Examples

¹ Schriftliche Fassung des Vortrags gehalten auf der ALLC / ACH 2002 Universität Tübingen

0. Introduction

Wordclass tagging and lemmatization nowadays gains more and more importance¹. It is required for syntactic analyses in dialog systems, for prosody prediction in speech synthesis and for syntactic annotations of large text corpora.

There seems to be a need for systems which can be implemented as modules in other NLP-systems.

In the following there will be described a system for unrestricted German texts, but it has to be mentioned that the structure of this tagger can be adopted to languages similar to German, that means to languages with a comparably high rate of inflectional phenomena.

The system LEMMA3 presented here is based on the program system LEMMA2², written in the early eighties using the programming language PL/1 and designed for the use on main frames.

The results and the speed of LEMMA2 were fairly good in those times, but as we intended to use the program on pc's and because of the meanwhile mostly outdated language PL/I we decided to rebuild the system and to redesign it at a large scale. This redesign of the LEMMA2 algorithm resulted in a much faster system, which can be used in speech synthesis systems at nearly real time, that means, that it produces no remarkable delay, with a very efficient deflection-modul using

- an extendable deflection list,
- a disambiguation facility for homographs within the closed word classes, and
- a mostly stringent separation of the data and the algorithms.

C++ was chosen as programming language because of its easy portability to many other platforms.

2. Functionality

The following sample input and output shall give a first glimpse of the functionality of LEMMA3:

Input:			
Was aber den Tanz betraf, so meisterte Herr Knaak ihn womöglich in noch höherem Grade.			
Output:			
word	basic form	word class	
Was	wer,was	PR	
aber	aber	KO	
den	der,die,das	PR	
Tanz	Tanz	SU	
betraf	betreffen	VE	V3
so	so	AV	
meisterte	meistern	VE	V3
Herr	Herr	SU	
Knaak	Knaak	SU	
ihn	er,es	PE	
womöglich	womöglich	AV	
in	in	PP	
noch	noch	AV	
höherem	hoch	AD	
Grade	Grad	SU	

As one can see, to each word form a basic form is added together with the indication of the word class and – in the case of verbs – of the inflection.

The codes will be explained at the end of the paper.

The program processes text in three steps:

- **Dictionary Lookup:** Every text word form is compared with a dictionary containing the – at this stage still partly homographic – elements of the closed word classes
- **Deflection:** Word forms belonging to the open, inflected word classes are processed by the modul MORPH
- **Disambiguation:** The homographs –the ones from the closed word classes and the not yet completely resolved inflection indications of verb forms –are disambiguated by the modul UMGEBUNG

LEMMA3 detects the following word classes based on the classic scheme of word classes³:

verb
 noun
 adjective
 adverb
 pronoun (incl. article)
 conjunction
 preposition
 postposition
 numeral
 interjection

and as a quasi word class:

verbzusatz (sense changing seperable morphs (prefixes)
added to verbs like the morph '**an-**' in '**ankommen**'
<**to arrive**> vs. '**kommen**' <**to come**>; compare
for example '**er kommt an**' <**he arrives**>:
then the word form '**an**' gets this word class
assigned)

In the case of abbreviations, which cannot be resolved unambigously, like '**u.**',
which can be used for '**und**' <**and**> as well as for '**unter**' <**under**>, only
'abbreviation' is indicated as word class.

3. Structure

The algorithm is mainly word form based, except for the third module UMGEBUNG, which uses some sentence context in order to disambiguate homographs, the rules of which are part of the program and not stored in an external data set.

The first step after the tokenization is the **dictionary look-up**.

The dictionary used is a full form dictionary containing the elements of the closed word classes. The information assigned to a match consists of the appropriate basic form and the word class indication, the latter possibly – if necessary – a homograph class, which will be disambiguated later on.

Each dictionary entry has the following structure:

keyword – base form – <word class>
<u>sample entries:</u>
denn denn <aK>
in in <PP>
seitlich seitlich <p9>

The homograph class '**aK**' (homograph between a conjunction and an adverb) is assigned to 'denn' and is to be disambiguated later on.

The names of the homograph classes have been chosen freely, the names mostly don't bear any special meaning.

The wordclass '**PP**' (preposition) is assigned to 'in'.

The homograph class '**p9**' (homograph between an adverb and a preposition) is assigned to 'seitlich' and is to be disambiguated later on.

The deflection for verbs, adjectives, and nouns is made by the second modul **MORPH**, which determines the word class and generates the proper basic form to a given word form.

This is done by means of a deflection list, which contains word stems or parts of them together with the allowed endings for a given basic form and its word class. The latter information can be taken only, if the tested word form ends in a valid combination of stem and ending.

For verb forms the inflection codes are supplied as well (at this point often still provisional).

The entries for nouns can be searched only, if a word form has an initial capital.

The distinction between adjectives and verbs with similar stem (for example 'lieb' vs. 'lieben') can be done only, if the endings are different; otherwise this distinction has to be done by the modul UMGEBUNG according to the syntactic context.

The structure of the deflection list is as follows:

key string (stem) – word class – basic form – <list of endings>

Some sample entries:

einzig AD einzig e, em, en, es, er, \$,
 geh VE gehen e,1 \$,11 st,3 est,12 (...)
 liebe SU liebe \$,

'\$' as an ending means 'no ending', the endings for verb forms include numeric codes for the inflection characteristics.

If the final string of the word form in question matches a list entry concatenated with one of the allowed endings given, then the word class is taken from the entry (in case of nouns after a successful check of an initial capital of the word form), the basic form of the current word is formed by substituting the matching string by the given basic form, in case of verbs the appropriate inflection codes are evaluated.

The following example may show, how the list is used:

word form in question: **'einziger'**
 successful match: **'einzig' + 'er'**
 word class added: **adjective ('AD')**
 generated basic form: **'einzig'**

For the explanation of some inflection codes for verbs the deflection list entry **'geh'** will be used:

deflection list entry:
 geh VE gehen e,1 \$,11 st,3 est,12 (...)

gehe = geh + e,1
1 = code **'G0'**: provisional code, homograph between 1. person singular present tense indicative active and 3. person singular present tense subjunctive active

gehst = geh + st,3
3 = code **'G2'**: 2. person singular present tense indicative active

geh = geh + \$,11
11 = code **'IS'**: imperative singular

The modul **UMGEBUNG** is the final component of LEMMA3 operating on whole sentences.

In a first step all word class assignments not yet done by the modul MORPH are added by rules, according to the word classes already determined for the surrounding word forms.

For example, if an undetermined word form with no initial capital ending in **'-e'** has a pronoun/article as predecessor and is followed by a noun, then the word class is **'adjective'**, the basic form is generated by removing the final **'-e'** from the word form.

if the predecessor is the personal pronoun '**ich**', then the word class is '**verb**', the basic form is generated by adding a final '**-n**' and the inflection is determined as 1.person singular present tense indicative active

Word form in question:
'heiße'

context a: diese heiße Tasse
PR ?? SU
'heiße' -> **AD**

context b: ich heiße Gerd
P1 ?? SU
'heiße' -> **VE**

PR means pronoun or article
SU means noun
P1 means personal pronoun '**ich**'
VE means verb
AD means adjective

As example for the disambiguation of the homographs found in the dictionary the rule for the homograph class '**p9**' will be described.

Rule for word forms marked as 'p9':

If the word following the word form in question is a pronoun, an article or a preposition, then it is a preposition;
else it is an adverb.

'seitlich' can be either a preposition or an adverb and is provisionally marked as '**p9**'.

Context 1: 'seitlich der Mauer'
?? PR SU
seitlich -> **PP**

Context 2: 'Er fuhr seitlich vorbei'
P3 VE ?? VX
seitlich -> **AV**

PR means pronoun or article
SU means noun
PP means preposition
P3 means personal pronoun '**er**'
VE means verb
VX means Verbzusatz
AV means adverb

The not yet completely resolved indications of the verbal inflection forms are disambiguated by context rules, which use the fact, that the 1. and 2. person of verbforms always must be accompanied by the corresponding personal pronoun. for example:

The word form '**gehe**' is assigned the word class **VERB** and the provisonal inflection code '**G0**' (because of the ending '**-e**')

Rule for the provisional code 'G0':

If an 'ich' is found in a defined context of the verb, '1. person singular present tense indicative active' (marked as **G1**) is assumed, otherwise '3. person singular present tense subjunctive active' (marked as **K3**).

Context a)	<i>'ich</i>	<i>gehe'</i>	
	P1	VE+??	
			-> gehe gehen VE G1
Context b)	<i>'er</i>	<i>gehe'</i>	
	P3	VE+??	
			-> gehe gehen VE K3

4. Limitations

The focus for the new conception of LEMMA3 had been laid on a clear manageable algorithm, the analysis data of which can be extended easily, more than on a system as linguistically sophisticated as possible.

This is reflected in the choice of the word classes and in the algorithm in general.

Some incorrect or fuzzy word class tags are produced systematically:

- **unresolved homographs treated as non-homographs:** As LEMMA3 has to add unambiguously all classifications and other informations, sometimes this can effect some results, which are not totally correct.
Apart from homographic forms like '**gehe**', which can partly be resolved by the aid of context rules, there are unresolvable homographs, for which there exist no rules of the type used in LEMMA3, for example the word form '**gerecht**', which can be either the adjective '**gerecht** <fair>' or the past participle of the verb '**rechen** <to rake>'. LEMMA3 has to make a clear choice; in this case Lemma3 always will treat the word form in question as an adjective.
- **Word forms with an initial capital**, which are not nouns in their context, (at the beginning of a sentence), but which resemble nouns, are treated as nouns. For example in '**Braten Sie Fisch?**' (**Do you roast fish?**) '**Braten**', written with initial capital because of its position at the beginning of a sentence always is treated as noun (**der Braten** <the roast>).
- **Word forms with an identical word class, which may belong to different basic forms:** There has to be made a choice as well (for example '**Summen**' nominative singular from '**das Summen** <the humming>' or plural from '**die Summe** <the sum>') is assigned the second basic form only.
Another example: the verb form '**du saust**' can be derived either from the verb '**sausen** (to rush)' or from the verb '**sauen** (to make a mess)'. The first basic form will be chosen always.

All those decisions had to be made when compiling the analysis lists.

Fuzzy assignemnts

As for the indication of the verbal inflection there are two major restrictions:

- The **infinitve** formally cannot be distinguished from the **3rd person plural present tense indicative active**, so the infinitive always has the inflection indicator 'G6', that means 3rd person plural present tense indicative active. The final decision then has to be made by the human user.
- **Forms of the compound tenses** (like '**ich werde gehen**') can be analysed only as seperated parts ('**werde**' as 1. pers. sg. present tense and '**gehen**' as infinitive).

As for the other word classes there are the following restrictions:

- **Adjectives used as adverbs:** Some assignments are homographic: The word class **adjective** often may be assigned to adjectives used syntactically as adverbs. (for example **'er singt gut'**)
- Some **homographs between adverb and verbzusatz** only can be resolved semantically, which is not possible with the LEMMA4-algorithm, and therefore are assigned the homograph word class **'AX'** (for example **abwärts**)
- If the word class **past participle** (marked as **'VA'**) **is assigned together with an inflection indication ('verbrachten' past participle of the basic form 'verbringen' and the inflection indication 3rd person plural past tense indicative active)**, this means that the verb form can be either an inflected form or a past participle, depending on the context, which has to be resolved by a human.

<p>compare the two sentences</p> <ul style="list-style-type: none">- sie verbrachten die Ferien zuhause<they spent their holidays at home> <p>and</p> <ul style="list-style-type: none">- die verbrachten Tage<the spent days>.

- **Non-nouns, which are used as nouns**, for example adjectives in **'der Gute kam ins Haus'** **<the good one entered the house>** are classified as if not capitalized. This is done because of the general capitalization of the first word form in a sentence.

5. Evaluation

A first evaluation has been made using the first two chapters of Thomas Mann's 'Tonio Kröger'. We have chosen this author because of his rich use of tenses, moods and inflection forms. Other texts will follow.

As 'error' will be understood a wrong assignment within the limits of the LEMMA3-algorithm including the incorrectly made assignments as described in the paragraph 'limitations'.

The text sample consisted of about 5600 running words.

- The total error rate found was **0.71 %**.

If differentiated between incorrect word class assignments and incorrect basic forms, but correct word class, the rates were as follows:

- wrong word class assignment: **0.48 %**
- wrong basic form only, but
correct word class: **0.23 %**

These results can be regarded as fairly good, if one considers the complex structure of Thomas Mann's texts.

6. Intended Use of LEMMA3

At our institute there has been a need for a fast word class tagger for speech synthesis purposes as well as for corpus linguistics.

The first implementation of LEMMA3 as a modul will be in the BOSS-System⁴ (**B**onn **O**pen **S**ynthesis **S**ystem), for which purpose real time processing is needed, and in RETIVOX⁵, a text-to-speech application, which provides an interface between telephone and e-mail; moreover it will be used stand-alone in corpuslinguistical research and teaching at our institute.

LEMMA3 will be freely available for scientific purposes. Requests may be sent to willee@uni-bonn.de.

7. Examples

First the output example of the beginning will be explained.

Sample Output of the System LEMMA3

<u>word</u>	<u>basic form</u>	<u>word class</u>	
Was	wer,was	PR	
aber	aber	KO	
den	der, die, das	PR	
Tanz	Tanz	SU	
betraf	betreffen	VE	V3
so	so	AV	
meisterte	meistern	VE	V3
Herr	Herr	SU	
Knaak	Knaak	SU	
ihn	er, es	PE	
womöglich	womöglich	AV	
in	in	PP	
noch	noch	AV	
höherem	hoch	AD	
Grade	Grad	SU	

Word classes

AD – adjective

AV – adverb

KO – conjunction

PE – pers. pronoun

PP – preposition

PR – pronoun/article

SU – noun

VE – verb

Verbal Inflection

V3 – 3. person singular

past tense indicative active

Two additional examples from 'Tonio Kröger':

Input text 1:

Da er daheim seine Zeit vertrat, beim Unterricht langsamen und abgewandten Geistes war und bei den Lehrern schlecht angeschrieben stand, so brachte er beständig die erbärmlichsten Zensuren nach Hause, worüber sein Vater, ein langer, sorgfältig gekleideter Herr mit sinnenden blauen Augen, der immer eine Feldblume im Knopfloch trug, sich sehr erzürnt und bekümmert zeigte.

Result after processing by LEMMA3

<u>word form</u>	<u>basic form</u>	<u>word class</u>	<u>inflection</u>	<u>word form</u>	<u>basic form</u>	<u>word class</u>	<u>inflection</u>
Da	da	KO		nach	nach	PP	
er	er	P3		Hause	haus	SU	,
daheim	daheim	AV		worüber	worüber	KO	
seine	sein	PR		sein	sein	PR	
Zeit	zeit	SU		Vater	vater	SU	,
vertrat	vertreten	VE	V3 ,	ein	ein	PR	
beim	bei	PP		langer	lang	AD	,
Unterricht	unterricht	SU		sorgfältig	sorgfältig	AD	
langsamen	langsam	AD		gekleideter	kleiden	VA	
und	und	KO		Herr	herr	SU	
abgewandten	abwenden	VA		mit	mit	PP	
Geistes	geist	SU		sinnenden	sinnen	VP	
war	sein	VE	V3	blauen	blau	AD	
und	und	KO		Augen	auge	SU	,
bei	bei	PP		der	der,die,das	PR	
den	der,die,das	PR		immer	immer	AV	
Lehrern	lehrer	SU		eine	ein	PR	
schlecht	schlecht	AD		Feldblume	feldblume	SU	
angeschrieben	anschreiben	VA		im	in	PP	
stand	stehen	VE	V3 ,	Knopfloch	knopfloch	SU	
so	so	AV		trug	tragen	VE	V3 ,
brachte	bringen	VE	V3	sich	er,sie,es	PE	
er	er	P3		sehr	sehr	AV	
beständig	beständig	AD		erzürnt	erzürnen	VA	G3
die	der,die,das	PR		und	und	KO	
erbärmlichsten	erbärmlich	AD		bekümmert	bekümmern	VA	G3
Zensuren	zensur	SU		zeigte	zeigen	VE	V3 .

Input text 2:

Denn es war das Merkwürdige, daß Tonio, der Hans Hansen doch um seine Daseinsart beneidete, beständig trachtete, ihn zu seiner eigenen herüberzuziehen, was höchstens auf Augenblicke und auch dann nur scheinbar gelingen konnte.

Result after processing by LEMMA3

<u>word form</u>	<u>basic form</u>	<u>word class</u>	<u>inflection</u>
Denn	denn	KO	
es	es	P3	
war	sein	VE	V3
das	der,die,das	PR	
Merkwürdige	merkwürdig	AD	,
daß	daß	KO	
Tonio	tonio	SU	,
der	der,die,das	PR	
Hans	hans	SU	
Hansen	hansen	SU	
doch	doch	AV	
um	um	PP	
seine	sein	PR	
Daseinsart	daseinsart	SU	
beneidete	beneiden	VA	V3 ,
beständig	beständig	AD	
trachtete	trachten	VE	V3 ,

<u>word form</u>	<u>basic form</u>	<u>word class</u>	<u>inflection</u>
ihn	er,es	PE	
zu	zu	PP	
seiner	sein	PR	
eigenen	eigen	AD	
herüberzuziehen	herüberziehen	VE	G6 ,
was	wer,was	PR	
höchstens	höchstens	AV	
auf	auf	PP	
Augenblicke	augenblick	SU	
und	und	KO	
auch	auch	AV	
dann	dann	AV	
nur	nur	AV	
scheinbar	scheinbar	AD	
gelingen	gelingen	VE	G6
konnte	können	VE	V3 .

**Word class codes produced by LEMMA3
(short list)**

- VE** – all verb forms except for:
VP – present participle
VA – past participle
- SU** – nouns
AD – adjectives
AV – adverbs
VX – verbzusatz
AX – homograph AD–VX and AV–VX
[resp.]
- PR** – pronouns (incl.articles, excl.pers.
[pronouns])
- personal pronouns:**
P1(1.sg) – 'ich'
P2(2.sg) – 'du'
P3(3.sg) – 'er, sie, es'
P4(1.pl) – 'wir'
P5(2.pl) – 'ihr'
PE – all other pers. pronouns
- KO** – conjunction
NU – numerals
PP – prepositions
PO – postpositions

- IJ** – interjections
AK – abbreviations

**Marks of the verbal Inflection
(active only)**

- G1 – G6**:1. pers. sing. to 3. pers. plural
present tense indicative
- V1 – V6**:1. pers. sing. to 3. pers. plural
past tense indicative
- K1 – K6**:1. pers. sing. to 3. pers. plural
subjunctive (I or II)
- IS**: imperative singular
IP: imperative plural

¹ cf. Hans van Halteren (1999), *Syntactic Wordclass Tagging*,
Dordrecht/Boston/London

² see among others: Gerd Willée (1982), *Das Programmsystem LEMMA2 - eine Weiterentwicklung von LEMMA*. IKP-Arbeitsbericht (Abt. LDV) Nr. 2, Bonn (published as manuscript)

and: Gerd Willée (1987), *Lemmatisierungsprobleme am Beispiel des Deutschen*. Vortrag, gehalten auf dem 35. Kolloquium über die Anwendung der EDV in den Geisteswissenschaften in Tübingen

³ The system of word classes used in LEMMA3 corresponds to the classic system, completed by the (quasi-)word class 'Verbzusatz' (separable semantically changing prefixes of German verbs, e.g. 'um-', 'ab-', 'zurück-').

⁴ see among others: Stöber, Karlheinz (2001), "Bestimmung und Auswahl von Zeitbereichseinheiten für die konkatenative Sprachsynthese", Doctoral Thesis, University of Bonn, Germany

Klabbers E.A.M., Stöber K., Veldhuis R., Wagner P., Breuer S. (2001), "Speech synthesis development made easy: The Bonn Open Synthesis System", In: Proc. Eurospeech'2001, Aalborg, Denmark, Vol. 1, pp. 521-524

⁵ RETIVOX is being developed by Dr. B. Schröder and Dr. Th. Portele, IKP, University of Bonn (e-mail: B.Schroeder@uni-bonn.de).