DeLex, a freely-avaible, large-scale and linguistically grounded morphological lexicon for German

Benoît Sagot

DeLex, a freely-avaible, large-scale and linguistically grounded morphological lexicon for German

Benoît Sagot

Alpage, INRIA & Université Paris Diderot, bâtiment Olympe de Gouges, 75013 Paris, France
benoit.sagot@inria.fr

Abstract

We introduce DeLex, a freely-avaible, large-scale and linguistically grounded morphological lexicon for German developed within the Alexina framework. We extracted lexical information from the German wiktionary and developed a morphological inflection grammar for German, based on a linguistically sound model of inflectional morphology. Although the developement of DeLex involved some manual work, we show that is represents a good tradeoff between development cost, lexical coverage and resource accuracy.

Keywords: Morphological Lexicon, German Morphology, Lexicon Development

1. Introduction

Contrarily to, e.g., syntax or semantics, inflectional morphology involves finite datasets over which acceptability judgments are most often clear-cut. It is no surprise that morphology is therefore one of the key areas where collaborative work is being carried out, which involves computational linguistics, formal linguistics, linguistic typology, and descriptive linguistics. This constitutes one of the main objectives of recent work for renewing the formalization and implementation of inflectional morphology within Alexina, an NLP framework for developing and encoding morphological and syntactic lexicons (Sagot, 2010), whose initial morphological formalism is described in (Sagot, 2005).[1] In particular, we have adapted and integrated within Alexina a formal model of inflectional morphology, named ParSTU (Walther, 2011; Walther, 2013b), which we shall sketch below. The result is both an extension of Sagot's (2005) morphological formalism and an implementation of ParSTU. It comes with various companion tools dedicated to quantitative morphological analysis. This new version of Alexina's morphological layer, described and named Alexina$_{\text{ParSTU}}$ in (Sagot and Walther, 2013), has been mostly used until now for acquiring new insights into specific linguistic issues within morphology or at the iterface between morphology and other components of grammar (Sagot and Walther, 2013; Walther, 2013a). In this paper, we want to show how this morphologically and typologically motivated framework can improve the theoretical soundness and the pratical maintainability of a morphological lexicon and its associated morphological description (or morphological grammar), as well as the speed of the development of such a language resource.

In this paper, we describe the recent development of DeLex, a new Alexina morphological lexicon for German. Apart from DeLex, and to our best knowledge, there has been surprisingly no freely available morphological lexicon for German until recently, as pointed out by Adolphs (2008). The only available lexicon is apparently the medium-scale lexicon distributed with the `morphisto` morphological analyzed (Zielinski and Simon, 2009) (18,624 entries among which 17,749 "base stems", i.e., mostly lemmas but also irregular inflected forms).[2,3] Moreover, German morphology is not strictly concatenative. It involves various kinds of *non-canonical* phenomena, in the sense of (Corbett, 2003), three of them being particularly widespread: stem alternation, syncretism and overabundance (in the sense of (Thornton, 2011), see below). We shall come back to this below in more detail, but this overall picture makes it the ideal testbed for validating Alexina's new morphological formalism through the development of a new large-scale resource: ParSTU, the underlying formal model, was specifically designed to address non-canonical phenomena. Our implementation within Alexina is now compatible with non-concatenative phenomena. And we believe the resulting resource is the first freely-available large-scale morphological lexicon for German.

2. A few words on the Alexina$_{\text{ParSTU}}$ model of inflectional morphology

As mentioned above, the morphological level of the Alexina formalism has recently been the focus of important collaborative efforts between computational and formal morphologists (Sagot and Walther, 2011; Sagot and Walther, 2013), resulting in the adaptation and implementation of the ParSTU model of inflectional morphology (Walther, 2011; Walther, 2013b), while retaining most features from the existing morphological framework (Sagot, 2005), thus resulting into Alexina$_{\text{ParSTU}}$. New Alexina$_{\text{ParSTU}}$ features include, among others:

- an arbitrary number of realisational levels (e.g., stem level, thematic level, first exponence level,

[1] Alexina lexicons have been developed for a fair amount of languages, among which French (the Le*fff*), Spanish (the Le*ff*e), Polish, Slovak, Persian, Kurdish or English. Other Alexina resources have been created by adapting existing freely-available resources. More recently, partial resources covering only (part of) one category have been developed for quantitative linguistics purposes, e.g., for Latin, Maltese or Khaling (Sino-Tibetan).

[2] Distributed under the Creative Commons 3.0 BY-SA licence.

[3] For instance, the free dictionary in the Unitex platform (Paumier, 2003), contains 300,000 word forms that constitue 10% of the CISLEX lexicon (Langer et al., 1996).

second exponence level); realisation rules, which are clustered in *zones* or in *tables* are level-specific (e.g., a German adjective could be modeled as having a non-marked, a comparative and a superlative stem which use the same exponence zone as the base stem as far as gender/number/case suffixes are concerned);

- realisation rules within a zone or table can be structured in *blocks* in the sense of (Stump, 2001);
- generic morphological operations that can be defined in a morphological grammar, which allow for modeling vowel alternations, reduplications, and other non-concatenative operations;
- mechanisms for modeling stem alternations, including suppletive stems;
- a mechanism for modeling form suppletion;
- structured morphological tags (flat morphological feature structures) and support of feature structure unification throughout morphological descriptions;
- lexical entries can specify explicitely information about deficiency (i.e., feature structures for which they do not have inflected forms, contrarily to most entries for the same category).

3. A few words on German morphology

German morphology is not stricly concatenative, in particular because nominal and verbal inflection involves vowel alternations (*ablaut* and *umlaut*) at the stem level, leading to stem allomorphy. (e.g., *Baum* 'tree', *Bäume* 'trees', or *fahren* 'drive', *(er) fährt* '(he) drives', *(er) fuhr* '(he) drove'). In addition, overabundance is massive, in particular within nominal and adjectival paradigms, both at the stem and at the exponence levels. For instance, at the exponence level, many masculine and neuter nouns can bear the suffix *-s* or *-es* for the GEN.SG, and/or the null suffix or the suffix *-e* for the DAT.SG (e.g., *Mann(e)s* 'man$_{\text{gen.sg}}$', *Mann(e)* 'man$_{\text{dat.sg}}$').

The stem level also shows overabundance for these nouns, specifically for the plural stem. As a result, the manual development of a morphological grammar for German was simplified and speeded up thanks to notions defined in PARSLI and implemented in AlexinaPARSLI. For example, our morphological grammar involves two realisational levels for adjectives and nouns, namely one stem level and one exponence level, as mentioned above. Indeed, all variation within adjectival inflection lies at the stem level, i.e., in the way the comparative and superlative stems are built. It can resort to morphonology and/or morphological irregularities, such as stem suppletion (*gut* 'good', *besser* 'better' *best-* 'best'), stem-related deficiency (*alkoholfrei* 'alcool-free' has no comparative or superlative) or stem overabundance (*frei* 'free', *freier* 'freer', *freist-* or *freiest-* 'freest'). Verbs involve an additional exponence level, which uses the unique adjectival exponence zone for inflecting the past participle. Note that the way this exponence zone is defined and used by past participles prevents them from having comparative and superlative forms.

In addition, the interface between morphology, syntax and the lexicon involves at least two massive and problematic phenomena: compounding and verbs with separable particles. Compounding is a massive phenomenon in German, particularly for nouns (e.g., *Tageslicht* 'day light'). Our goal being to create an inflectional lexicon, constuctional processes need not be modeled in DeLex. However, one could consider modeling compound components (e.g., *licht*) as belonging to the set of forms associated with their corresponding lexical entries, leaving composition rules to another level of morphological modeling. On the other hand, many compounds are lexicalized, or at involve at least strong conventional preferences, thus suggesting that being the result of a morphological process and being a reasonable lexical entry should be considered as two distinct properties (Gaeta and Ricca, 2009). In this work, we decided to leave these difficult theoretical issues aside and rely on a purely empirical strategy: we shall include in DeLex all lexical entries found in the German Wiktionary, be they compounds or not. Moreover, our inflectional grammar will not generate compound components, as we consider that compoud creation rules belong to another layer of morphological modeling, which should be able to take as an input inflectional lexical entries (be it in the form of a stem or of an inflected form). In other words, DeLex does not provide any direct and dedicated way to analyse compounds it does not already contain, but could be used as a source of lexical information for developing a compound analysis module.

The last phenomenon we shall briefly discuss is that of verbs with a separable (prefixal) particle. Given a simple verbal entry, such as *geben* 'give', one can find derivationaly related entries that also include one (sometimes several) prefixal elements, which can be of various kinds. Among them, the most challenging are *separable particles*. For example, *aufgeben* 'give up' involves a separable particle, *auf*, which is can be associated with the preposition *auf*.[4] It is said to be separable because some cells in *aufgeben*'s paradigm contain a form without that particle (e.g., *gebe* 'I give (up)') but require it to be at a well-defined other place in the clause (*Ich gebe niemals auf* 'I never give up'). Prefixal realization rules at work in other cells insert the corresponding morph between the separable particle and the stem (e.g., past participle *aufgegeben*). In this work, we consider the result of all realisation rules that are applied to the stem involved — including the extraction of the separable particle as an independent "word" —, together with an information associated with the inflected entry that make explicit the fact that this particle is to be found elsewhere in the clause, as an inflected form of a verb with a separable particle. In other words, both *aufgegeben* and *gebe* will be inflected

[4] Not only preposition-like elements can serve this role, see for example a noun-like element in *heimkommen* 'come home' (*Heim* 'home' [noun] + *kommen* 'come') or an adjective-like element in *freigeben* 'set free, release' (*frei* 'free' [adj.] + *geben* 'give'). Such elements are virtually always separable.

entries for the lexical entry *aufgeben* in DeLex, but the latter will also specify the need for an *auf* elsewhere in the clause.[5,6]

4. Developing DeLex

The efficient development of this morphological grammar has been realised together with the extraction of lexical data and continuous validation of both the grammar and the lexicon *via* the paradigms they generate. More precisely, it can be described as a three-step process, starting with the German Wiktionary.[7]

First, we have developed a dedicated rule-based tool that automatically extracts German lexical entries from a dump of the German Wiktionary, which can be considered as a semi-structured, noisy and incomplete lexical dataset.[8] We limited the use of this tool to nominal, verbal, adjectival and adverbial lexical entries.[9] For each of them, we extracted the citation form as well as the partial morphological paradigms that can be found within Wiktionary articles (see Figure 1). We converted this into triples of the form (citation form,[10] inflected form, morphological tag), in which morphogical tags are manually specified based on Wiktionary cell names (e.g., "Befehl_du" is converted into IMP.SG). In some cases, including for the case of the example shown in Figure 1, the partial paradigms we extracted exhibit overabundance: for some cells within the paradigm, such as imperative singular (IMP.SG) in this case, two distinct forms are possible (*mach* and *mache*, 'do'). In the German Wiktionary, this overabundance can be expressed using brackets (*mach(e)*), explicitely within one cell (*geh!, gehe!* 'go'), or as belonging to two distinct parts of the paradigm that realize the same set of feature structures (e.g., when a noun has two distinct plural stems, both stems lead to a distinct plural subparadigm).

In a second step, and for each of the four open-class categories, we applied an algorithm that automatically builds inflection classes to the set of triples extracted during the first step.[11] The underlying objective is not to build DeLex's final inflection classes (recall that the triples extracted during the first step only cover parts of the paradigms). Rather, it aims at identifying sets of lexical entries that share a common inflectional behavior. This automatic inflection class construction algorithm works as follows. First, it performs on each partial paradigm (i.e., on each set of triples that share the same part-of-speech and the same citation form) the following actions:

- it identifies the longest initial substring of the citation form (*machen*, in our example) that is also a (non-necessarily initial) substring of all inflected forms of the (partial) paradigm (*mach* in our example);
- it considers this substring as the stem;[12]
- it considers that, for each inflected for in the (partial) paradigm, the initial substring to the left, resp. right, of the stem is its prefix, resp. suffix, and associates the resulting (prefix, suffix) pair to the triple's morphological tag;
- it sorts (prefix, suffix) pairs according to the corresponding morphological tag (alphabetically);
- it considers the result of this sort as a signature of the inflection class that has generated the (partial) paradigm at hand, i.e., the inflection class for the corresponding lexical entry.

Once this has been achieved on all (partial) paradigms, we generate a morphological description (i.e., morphological grammar) that defines all extracted inflection classes, provided they are associated with at least 3 lexical entries. Indeed, inflection class signatures contain all the information needed for defining the inflection classes' realization rules. Inflected forms for lexical entries requiring inflection classes used by less than 3 lexical entries are listed explicitely and are given the special inflection class 0 which generates no (additional) inflected form.

For the third step of DeLex's construction, we performed two costly manual task,[13] namely:

- the development of a complete morphological description of German, which is not restricted to the partial paradigms extracted from the German wiktionary, and which is linguistically motivated, based on the Alexina$_{\text{PARSLI}}$ formalism mentioned above. This description relies for example on an explicit modeling

[5] In fact, the situation is even more complex. A verb such as *rollerbladen* behaves like a simple verb as far as particle separation is concerned (*Ich rollerblade* 'I rollerblade') but as a verb with a separable particle when inflectional prefixation is involved (past participle: *rollergebladet*). This is correctly modeled in DeLex.

[6] The other major class of prefixal elements is that of non-separable particles/prefixes, which prevent any inflectional prefixation to apply: compare *geben* 'give' and *gegeben* 'given' with *vergeben* 'forgive' and *vergeben* 'forgiven.'

[7] `http://de.wiktionary.com`

[8] This extraction task is non-trivial as each page covers one citation form, which can represent several lexemes (i.e., lexical entries) in various languages (not only German).

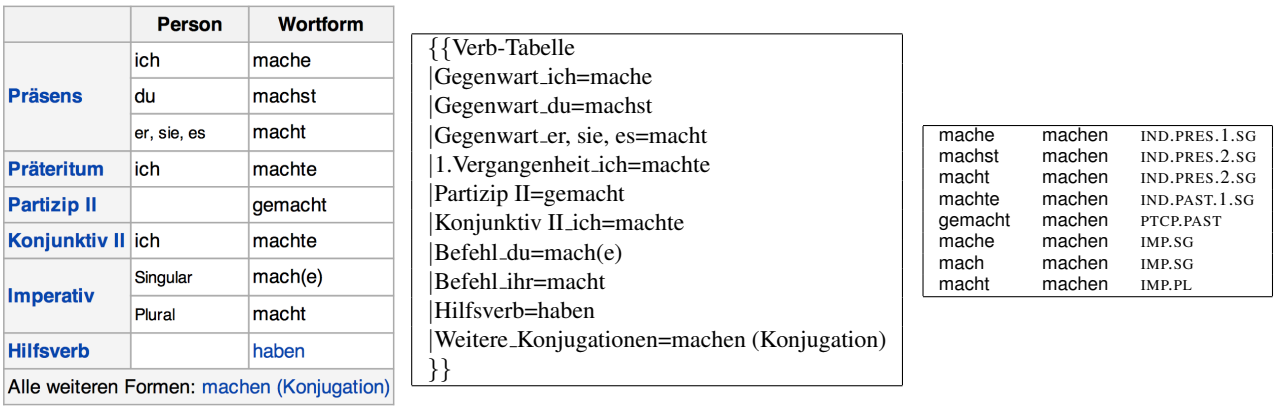
[9] Inventories of closed class lexical entries (determiners, pronouns, etc.) have been extracted directly from the TIGER treebank (Brants et al., 2002; Smith, 2003), ignoring hapaxes, then adapted and extended manually. We do not provide details in this abstract for space reasons, and focus on the Wiktionary-based extraction of open-class lexical entries.

[10] "Citation form" is more correct than "lemma", as a lemma is rather an equivalence class of inflected forms, whereas the citation form here is one of the inflected forms of a lemma, chosen as its representative.

[11] We use the term "inflection class" with its usual sense here. In PARSLI, the adequate term would be here "inflection pattern."

[12] As said, we try and identify sets of lexical entries that share a common inflectional behavior. Clearly, lexical entries involving more than one stem will be treated suboptimally. This will be fixed later on by manually defining stem realization classes and merging sets of lexical entries that have the same behavior at the exponence level.

[13] A few dozens of hours for a non-native speaker of German with a good linguistic knowledge of this language and a B1/B2 proficiency level, with access to online dictionaries and other online resources.



	Person	Wortform
Präsens	ich	mache
	du	machst
	er, sie, es	macht
Präteritum	ich	machte
Partizip II		gemacht
Konjunktiv II	ich	machte
Imperativ	Singular	mach(e)
	Plural	macht
Hilfsverb		haben
Alle weiteren Formen: machen (Konjugation)		

Excerpt of the page for *machen* 'do' in the German Wiktionary

```
{{Verb-Tabelle
|Gegenwart_ich=mache
|Gegenwart_du=machst
|Gegenwart_er, sie, es=macht
|1.Vergangenheit_ich=machte
|Partizip II=gemacht
|Konjunktiv II_ich=machte
|Befehl_du=mach(e)
|Befehl_ihr=macht
|Hilfsverb=haben
|Weitere_Konjugationen=machen (Konjugation)
}}
```

Corresponding source code (wiki syntax)

mache	machen	IND.PRES.1.SG
machst	machen	IND.PRES.2.SG
macht	machen	IND.PRES.2.SG
machte	machen	IND.PAST.1.SG
gemacht	machen	PTCP.PAST
mache	machen	IMP.SG
mach	machen	IMP.SG
macht	machen	IMP.PL

Automatically extracted partial paradigm

Figure 1: Illustrating the first step of DeLex's construction on an example: extracting lexical entries and partial paradigms from the German Wiktionary

of vowel alternations within verbal and nominal steps; it allows for explicitely specifying suppletive stems within the lexicon, which is for instance necessary for several comparatives and superlatives and for some verbs; it also contains constraints associated with inflection classes and with rules, in order to restrict the use of an inflection class or realisation rules to compatible stems;

- the conversion of the lexicon extracted automatically in the previous steps, in order to generate a lexicon that relies on these manually defined inflection classes.

This allowed us to produce a lexicon which is satisfying both from a linguistic and from a coverage point of view. It also allowed us to identify and correct many errors within the German Wiktionary. This is the case for example whenever the inflected forms generated by our lexicon for a given lexical entry were different from what had been extracted from Wiktionary, and whenever some inflection classes of some realization rules could not be applied for a given cell and a given lexical entry because of the above-mentioned constraints.

The result of this three-step process is illustrated in Figure 4, which displays both the lexical entry for our running example, the verb *machen* 'do', together with excerpts of the morphological description that are relevant for inflecting this entry.

5. The resulting lexicon: DeLex

The resulting lexicon, which constitutes the first version of DeLex, contains 63,017 intensional entries (i.e., triples containing a citation form, a part-of-speech and an inflection class). More precisely, as far as open classes are concerned, there are 6,530 adjectival entries, 39,670 nominal entries, 4,899 verbal entries and 904 adverbial entries — see Figure 2 for a few examples.[14] Once inflected, these intensional entries generate over 2.3 million extensional entries, i.e., 4-tuples of the form (inflected form, citation form, part-of-speech, morphological tag) — see Figure 3 for a few examples. More precisely, the extensional (inflected) lexicon contains around 441,000 adjectival forms, 301,000 nominal forms, 486,000 verbal forms and 908 adverbial forms.

DeLex is already freely-available for download under the LGPL-LR licence, both in the form of an intensional lexicon (lexical entries and inflectional grammar) and in the form of an extensional lexicon (inventory of inflected forms associated with morphological information, directly usable in tools such as part-of-speech taggers, unknown word analysis modules for parsers, and many others).[15]

begeben	v-a-e-i:fs:noge:0
begegnen	v-std:noge
begehen	v-std:strong:fs:noge:0/,,beging,begang
begehren	v-std:fs:noge:0
begeistern	v-std:hs:noge
beginnen	v-a-o-i:fs:noge:0
begleichen	v-i:fs:noge

Figure 2: A few (simplified) verbal intentional entries

achtenswertestes	ADJ	super.sg.neu.nom.primary.long	achtenswert
achter	ADJ	plain.pl.gen.primary.short	achte
achter	ADJ	plain.sg.fem.dat.primary.short	achte
achter	ADJ	plain.sg.fem.gen.primary.short	achte
achter	ADJ	plain.sg.masc.nom.primary.short	achte
achterlich	ADJ	plain.noagr.long	achterlich

Figure 3: A few (simplified) adjectival extentional entries. In this figure, an extensional entry includes an inflected form, a category, a morphological tag, and the lemma's citation form and inflection class.

[14] One should note that, in German, adjectives can often be used as adverbs, which explains why there are only 904 adverbial entries: such entries only cover non-adjectival adverbs, such as *schon* 'aldready,' *zugleich* 'similarly,' *etwa* 'approximately, around,' and so on.

[15] At the URL `https://gforge.inria.fr/frs/?group_id=482`, one can find the `alexina-tools` package, which is required for inflecting DeLex (version 1.5 and higher is required), as well as the `delex` package itself, whose version 0.1 is described here. The `delex` package also contains the resulting extensional morphological lexicon, i.e., entries of the form (inflected form, category, citation form) (see file `delex.mlex`).

Lexical entry:

machen v-std:fs:ge:0

Explanation:

- "v-std" is the inflection pattern (or improperly the inflection class) of verbs that only have one stem for all cells (regular, most frequent case);
- "fs" specifies one of the possible behaviors of the schwa at the stem-suffix interface;
- "ge" specifies that the past participle is realized with the prefix *ge-*
- "0" specifies that, beyond the standard *-e* imperative singular form (*mache*), an alternative suffixless form est possible (*mach*).

Morphological description (or morphological grammar):

- Several morphological operations, which are not readily avaible in the formalism, are defined explicitely. For example, the operation remove_sp removes the separable particle from a verb, provided it has one (e.g., *aufmachen* → *machen*).
- Various morphographemic rules are defined, such as the rule given below, which indicates that a suffix starting with *t-*, when place at the right of an element ending in *-t*, leads to the insertion of an extra *-e-*.

```
<sandhi source="t_t" target="t_et"/>
```

- The inflection pattern below specifies at the level 1 how the stem is built for each cell within regular verbal paradigms, and at the level 2 how affixal exponents should be computed.

```
<pattern name="v-std" cat="v">
  <subpattern>
    <realzone level="1" table="v-stems"/>
    <realzone level="2" table="v-infl"/>
  </subpattern>
</pattern>
```

- The stem construction table v-stems is not very informative, as verbs which use it have only one stem for all cells.
- The exponence table v-infl is the same for all verbs, although a few rules only apply on strong verbs and other for other verbs. We provide below a simplified version of this table, which covers indicative present forms for all non-strong verbs whose stem does not end in *-s*, *-z*, *-ß* or *-x*. Note the first rule, which removes the separable particle when relevant (their pattern indicate spge or spnoge instead of the ge in the pattern mentioned above).

```
<table name="v-infl">
  <form block="1" operation="remove_sp" except="inf" var="spge|spnoge"/>
  <form block="4" suffix="" features="short" var="ns"/>
  <form block="4" suffix="" features="short" except="ind.pres.1.sg" var="hs"/>
  <form block="4" suffix="" features="short.ind.pres.2|short.ind.pres.3.sg" var="fs"/>
  <form block="4" suffix="e" features="ind.pres.1.sg.long" var="hs|ns"/>
  <form block="4" fail="1" var="hs|ns"/>
  <form block="4" fail="1" features="long.ind.pres.2.sg|long.ind.pres.3.sg" var="fs"/>
  <form block="4" suffix="e" features="long"/>
  <form block="4" fail="1"/>
  <form block="5" suffix="" features="1.sg"/>
  <form block="5" suffix="st" features="2.sg"/>
  <form block="5" suffix="t" features="ind.pres.3.sg"/>
  <form block="5" suffix="n" features="1.pl"/>
  <form block="5" suffix="t" features="2.pl"/>
  <form block="5" suffix="n" features="3.pl"/>
</table>
```

Overabundance — a frequent phenomenon in German — is dealt with using an *ad hoc* 'length' feature, which can take one of the following values: short and long. In any case, realization rules are tried in their order of appearance within each block, until one is found that can be applied or that explicits a failure (fail="1").

Figure 4: The lexical entry for *machen* 'do' in DeLex, together with excerpts of the morphological description relevant for building its full paradigm. Everything involving participial forms has been discarded for the sake of readability.

6. Evaluation of DeLex

6.1. Recall on the TIGER corpus

A first way to evaluate DeLex is simply to assess its coverage on a gold annotated corpus. We have extracted all words from the TIGER corpus (Brants et al., 2002; Smith, 2003) and performed a look-up for each of them in DeLex. We have discarded all tokens involving only digits and punctuation signs and have allowed lowercase look-ups whenever case-sensitive loop-up failed. Because DeLex does not contain any named entities, the raw coverage figure computed this way, namely 88.9%, is a lower bound of DeLex's coverage. Unsurprisingly, the most frequent TIGER tokens unknown to DeLex are named entities.[16] On the other hand, ignoring all tokens involving at least one capital letter provides an upper bound of the coverage, as many tokens involved named entities will be discarded, but all common nouns as well (in German, common nouns are capitalized). This upper bound is as high as 97.8%. A maybe better estimate can be obtained by ignoring all tokens in TIGER that are annotated as "NE" (named entities). The resulting figure is 93.1%.

6.2. Comparison with the `morphisto` lexicon

Another way to evaluate DeLex is to compare it with the above-mentioned `morphisto` lexicon (Zielinski and Simon, 2009). The `morphisto` lexicon is an inventory of various lexical elements, among which "base stems." As far as nouns, adjectives and adverbs are concerned, almost all `morphisto` "base stems" are in fact citation forms in the same sense as for DeLex. Therefore, they can be compared with the one another.

DeLex contains 39,795 unique nominal citation forms, whereas `morphisto` contains 9,165 nominal "base stems." Among them, 6,476 are also DeLex citation forms, whereas 2,481 are not. The latter are mostly proper names (*Laos*, *Nike*), but also nouns that are indeed missing from DeLex. On the other hand, 33,319 DeLex nominal citation forms are not among `morphisto`'s "base stems."

DeLex contains 6,568 unique adjectival citation forms, whereas `morphisto` contains 3,212 adjectival "base stems." 2,123 are shared by both resources, 1,089 `morphisto` adjectival "base stems" are unknown to DeLex, and 4,445 DeLex citation forms are not `morphisto` "base stems."

Because the definition of what an adverb is or should be is not necessarily the same between DeLex and `morphisto`, we have looked for `morphisto` adverbial "base stems" within adverbial entries as well as within entries in all DeLex closed categories (i.e., excluding nouns, verbs and adjectives). Out of the 706 `morphisto` adverbial "base stems," 545 are known to DeLex as citation forms, whereas 161 are not.[17]

As far as verb forms are concerned, `morphisto` "base stems" are not citation forms any more. They are really stems, associated with an inflection class or a partial inflection class, or inflected forms associated with the corresponding morphological tag. We have therefore replaced all stems by one of the inflected forms that would be generated by the (partial) inflection class (e.g., we replaced the stem *brat* associated with the tag `VVPres1+Imp` by the form *brate*). Together with explicitely listed inflected forms from `morphisto`, we thus obtained a set of 4,243 unique verb forms. Out of these forms, 2,892 are known to DeLex as inflected forms. Out of the 1,041 other forms, the vast majority of them are infinitives, from rare verbs yet unknown to DeLex.

Overall, DeLex has therefore a much larger coverage than the `morphisto` lexicon, but there is still a significant room for improvement for DeLex. As mentioned in the conclusions of this paper, corpus-based extension of DeLex is among our first priorities for future work.

6.3. Lexicon-based improvement of a POS tagger

Another way to assess the quality and usefulness a lexicon is to perform a task-based evaluation. We have performed POS-tagging experiments for quantifying whether using DeLex as an external lexicon within a statistical POS-tagger leads to improvements over the same POS-tagger when only trained on the training corpus. For these experiments we have used MElt (Denis and Sagot, 2012), a statistical POS tagging system that relies on Maximum-Entropy Markov Models trained on an annotated corpus and optionaly an external lexicon. MElt has been trained on data for French, for which is is state-of-the-art, and several other languages. We have trained MElt on a slightly modified[18] version of the TIGER corpus. More precisely, the corpus we used contains 50,499 sentences (888,533 tokens), from which we selected the first 45,000 sentences (788,448 tokens) for training and the remaining 5,499 sentences (100,859 tokens) for evaluation. The unknown token rate in the evaluation subcorpus (rate of words in the evaluation subcorpus that do not occur in the training subcorpus) is 9.2%. Our modifed version of the corpus relies on a 65-tag tagset.

We have trained MElt on the training subcorpus, both without and with DeLex as an external lexicon (respectively $\text{MElt}_{\text{de}}^{\text{nolex}}$ and MElt_{de}).[19] Results, displayed in Table 1, show that using DeLex allows for a significant increase in POS tagging accuracy, in particular on unknown words.

7. Conclusion and perspectives

The next steps of DeLex's development shall consist in large-scale unknown word extraction from various corpora and (semi-)automatic lexical entry creation, following

[16] In decreasing frequency order, the three most frequent unknown tokens are *SPD* (party name), *Bonn* (city), *M.* (almost always part of *Frankfurt a. M.*, abbreviation for *Frankfurt am Main*, the city of Francfort). Note also some non-words such as *ap* or *rtr*, typical of article signatures.

[17] Apart from ADV (standard adverbs), the most frequent DeLex (i.e., TIGER) category for `morphisto` adverbial base stems are PTKVZ (separable verbal particle), PROROAV (pronominal adverb, e.g., *dafür* 'for that') and PROWAV (adverbial interrogative or relative pronoun, e.g., *warum* 'why').

[18] Our modification only consists in assigning to punctuation tokens a tag identical to the token itself, instead of one of the three punctuation tags.

[19] We extracted POS and morphological tags within DeLex and converted them into morphosyntactic categories compatible with our modified version of the TIGER corpus.

	TOKENS		ACCURACY	
	NB.	%	$MElt_{de}$	$MElt_{de}^{nolex}$
Known words	91,603	90.8%	97.8%	97.5%
Unknown words	9,256	9.2%	90.6%	88.8%
All words	100,859	100%	**97.2%**	96.7%

Table 1: Evaluation of the $MElt_{de}$ POS-tagger, which uses DeLex as an external lexicon, and comparison with its variant $MElt_{de}^{nolex}$ which does not use DeLex. The tagset involved contains 65 distinct categories.

previous comparable efforts on Spanish (Molinero et al., 2009) or French (Sagot et al., 2013). This should also allow us to automatically identify errors in the current version of DeLex, as non-covered forms of covered lemmas can only be the result or erroneous or missing lexical information and/or of errors in the inflectional grammar.

Apart from extending and improving DeLex, we have already planned or initiated work using DeLex for various purposes such as quantitative morphology, acquisition of derivational lexical information, advanced tokenization and statistical parsing.

Moreover, DeLex has already started to be used for extraction derivational morphlogical relations (Baranes and Sagot, 2014). Beyond derivation, dealing with the German constructional morphology in a linguistically sound and NLP-ready lexicon is a challenging task, that we would like to study in the future.

Acknowledgements

This work was partly funded by the French ANR national grant EDyLex (ANR-09-CORD-008) and the French "Investissements d'avenir" project PACTE.

8. References

Peter Adolphs. 2008. Acquiring a poor man's inflectional lexicon for German. In *Proceedings of LREC'08*, Marrakech, Morocco.

Marion Baranes and Benoît Sagot. 2014. A language-independent approach to extracting derivational relations from an inflectional lexicon. In *Proceedings of LREC'14*, Reykjavik, Iceland.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, pages 24–41.

Greville G. Corbett. 2003. Agreement: the range of the phenomenon and the principles of the surrey database of agreement. *Transactions of the philological society*, 101:155–202.

Pascal Denis and Benoît Sagot. 2012. Coupling an annotated corpus and a lexicon for state-of-the-art POS tagging. *Language Resources and Evaluation*, 46(4):721–736.

Livio Gaeta and Davide Ricca. 2009. Composita solvantur: Compounds as lexical units or morphological objects? *Rivista di Linguistica*, 1(21):35–70.

Stefan Langer, Petra Maier, and Jürgen Oesterle. 1996. CISLEX, an electronic dictionary for German. Its structure and a lexicographic application. In *Proceedings of COMPLEX 1996*, Budapest, Hungary.

Miguel Ángel Molinero, Benoît Sagot, and Lionel Nicolas. 2009. A morphological and syntactic wide-coverage lexicon for Spanish: The Le*ffe*. In *Proceedings of RANLP'09*, Borovets, Bulgarie.

Sébastien Paumier. 2003. *De la reconnaissance de formes linguistiques à l'analyse syntaxique*. Ph.D. thesis, Univerité de Marne-la-Vallée.

Benoît Sagot and Géraldine Walther. 2011. Non-canonical inflection : data, formalisation and complexity measures. In Cerstin Mahlow and Michael Piotrowski, editors, *Systems and Frameworks in Computational Morphology*, volume 100, pages 23–45, Zurich, Switzerland. Springer.

Benoît Sagot and Géraldine Walther. 2013. Implementing a formal model of inflectional morphology. In Cerstin Mahlow and Michael Piotrowski, editors, *Systems and Frameworks in Computational Morphology*, volume 380, pages 115–134, Berlin, Germany. Springer.

Benoît Sagot, Damien Nouvel, Virginie Mouilleron, and Marion Baranes. 2013. Extension dynamique de lexiques morphologiques pour le français à partir d'un flux textuel. In *Proceedings of TALN 2013*, pages 407–420, Les Sables d'Olonne, France, June.

Benoît Sagot. 2005. Automatic acquisition of a Slovak lexicon from a raw corpus. In *Lecture Notes in Artificial Intelligence 3658 ((c) Springer-Verlag), Proceedings of TSD'05*, pages 156–163, Karlovy Vary, Czech Republic.

Benoît Sagot. 2010. The Le*fff* , a freely available, accurate and large-coverage lexicon for French. In *Proceedings of the LREC'10*, Valletta, Malta.

George Smith. 2003. A brief introduction to the TIGER treebank, version 1. Technical report, Universität Potsdam.

Gregory T. Stump. 2001. *Inflectional Morphology. A Theory of Paradigm Structure*. Cambridge University Press, Cambridge, Royaume-Uni.

Anna M. Thornton. 2011. Overabundance (multiple forms realizing the same cell): A non-canonical phenomenon in italian verb morphology. In M. Goldbach M. Maiden, J. C. Smith and M.-O. Hinzelin, editors, *Morphological Autonomy: Perspectives From Romance Inflectional Morphology*. Oxford University Press.

Géraldine Walther. 2011. Measuring morphological canonicity. *Linguistica*, 51:157–180. Internal and External Boundaries of Morphology.

Géraldine Walther. 2013a. Controlling arbitrariness: descriptive economy as an index of inflectional complexity.

Géraldine Walther. 2013b. *Sur la canonicité en morphologie — Perspective empirique, formelle et computationnelle*. Ph.D. thesis, Université Paris-Diderot.

Andrea Zielinski and Christian Simon. 2009. Morphisto – an open source morphological analyzer for German. In *Post-proceedings of the FSMNLP'08*, pages 224–231, Amsterdam, The Netherlands. IOS Press.