Syncretism and How to Deal with it in a Morphological Analyzer: a German Example

Katina Bontcheva
Heinrich-Heine-University Düsseldorf
bontcheva@phil.uni-duesseldorf.de

Abstract

Syncretism is the area of the morphology-syntax interface where morphology fails the syntax. Inadequate treatment in the design of a morphological analyzer can lead to unbalanced performance of the analyzer either at generation, or at analysis. Furthermore, adequate and consistent treatment of syncretism is needed if the analyzer is to be used for language modeling, especially modeling of the syncretism. In this paper I will show that it is possible to create a morphological analyzer that can be tailored to various intended uses with minimal effort.

1 Introduction

Syncretism may seem to be a minor morphological phenomenon, but this is the place in the morphology-syntax interface where morphology fails syntax. Inadequate treatment in the design stage of a morphological analyzer can lead to undesirable ambiguities in generation or analysis.

A morphological analyzer can have different applications. It can be used in a pipeline with a POS tagger, a shallow or deep-syntactic parser, a semantic parser, for generation or language modeling, among other things.

Depending on the intended use, one might wish to avoid the ambiguity in analysis caused by multiple possible readings of syncretic forms if they are morphosyntactically fully specified. On the other hand, underspecification at the lexical level will lead to multiple output strings at generation.

In this paper I will show that it is possible to create a morphological analyzer that can be tailored with minimal effort to various intended uses.

In section 2 I will briefly discuss syncretism, what types of syncretism exist, and how one can model syncretism using or not using rules of referral. The prototypical finite-state morphological analyzer for German that I am currently working on will be described in section 3, while in section 4 I will present the paradigms of adjectival agreement in standard German that are heavily affected by syncretism. In section 5 I will explain on the basis of the German example and examples from other languages how with minimal changes one can tune the prototypical morphological analyzer to perform the different tasks outlined earlier in this section.

In section 6 I will draw some conclusions, and in the Appendix I will show a code excerpt.

2 Syncretism

Syncretism is the identity of two or more inflected forms of the same lexeme. The identity of two forms that belong to different lexemes should be treated as accidental homonymy. Thus the form *books* is not syncretic since *book-N.PL* and *book-V.PRES.3SG* belong to different lexemes. However, the form *book* is syncretic within the paradigm of the verb *book* since it is associated with a set of morphosyntactically distinct feature values, e.g., *book-V.PRES.1SG*, *book-V.PRES.2SG*. *book-V.PRES.1PL*. *book-V.PRES.2PL*, etc.

One of the characteristics of syncretism is *directionality*. "Directionality concerns the possible morphological affiliation of the syncretic form to one of its component values" (Baerman, Brown and Corbett 2005, p. 24).

Since syncretism involves a set of morphosyntactic values that are associated with a single form, the question is how exactly they are associated. There are two options (cf. Baerman, Brown and Corbett 2005, p. 133): a) the form is related to the set as a whole or b) the form is related to one of the values and the other morphosyntactic values "borrow" the form. Stump (2001) calls the former *symmetric rules* and the latter *directional rules*. Symmetric rules simply map a form/string to a set of values in one step, whereas directional rules entail more than one step. In the first step there is a mapping of a form/string to a particular value of the set, and in

Proceedings of the 11th International Conference on Finite State Methods and Natural Language Processing, pages 104–107,
St Andrews–Sctotland, July 15–17, 2013.

the consecutive step(s) this value is associated with the rest of the set. We call such directional rules *rules of referral* (cf. Zwicky 1985).

Lack of directionality is often caused by uninflectedness, loss of inflection, or the merger of the reflexes of two or more phonemes. Examples of directional and non-directional syncretism are presented in section 5.

Syncretism can be caused phonologically, i.e., it can be the result solely of a phonological rule, lexically, i.e., within a single lexical item, or morphologically, i.e., spanning over at least one inflectional class.

3 The Prototypical Morphological Analyzer for German

The prototypical morphological analyzer for German consists of a lexc lexicon that describes the morphotactics of the language, and of phonological and orthographical alternations and realizational rules, and possibly also rules of referral, that are handled elsewhere by finite-state replace rules.

The bases for the regular inflectional classes are stored separately in text files. Bases of words that are subject to morphographemic alternations (e.g. *Umlaut*) have abstract representations at the surface level and lemmata on the lexical level. There are no semantic features in the current version of the lexicon.

The tagset that is used in this version of the analyzer is compatible with the MULTEXT-East morphosyntactic specifications (cf. MULTEXT-East morphosyntactic specifications, Version 4, 2010). It was chosen in preference to the Stuttgart-Tübingen tagset (STTS) (Telljohann et al., 2009) because it implements atomic values and is compatible with the tagsets for other (European) languages.

Here is an excerpt from a text file that contains qualitative-adjective bases with *Umlaut*:

```
{alt}:{1lt}
```

On the left is the lemma (*alt* ‘old’) that will appear in the analysis output and on the right is the abstract form that contains the abstract symbol *1* for a lowercase *a* which is subject to *Umlaut* alternations under certain conditions.

And here is an excerpt from the lexc lexicon:

```
LEXICON Adjectives
LxAQUAL          Adj ;
LxAQUAL          AdjCmpSpl ;

LEXICON Adj
<"+A" 0:"+Uninfl"> # ;
<"+A" 0:"+Pos">  AStrong;
<"+A" 0:"+Pos">  AWeakMixed;

LEXICON AdjCmpSpl
<"+A" 0:"+Cmp" 0:"+Uninfl"> #;
<"+A" 0:"+Cmp">  AStrong;
<"+A" 0:"+Spl">  AStrong;
<"+A" 0:"+Cmp">  AWeakMixed;
<"+A" 0:"+Spl">  AWeakMixed;

LEXICON AStrong
…
```

This excerpt partially illustrates the morphotactics of the adjectives. The rest - inflection for gender/number/case - will not be presented because of space limitations. The excerpt shows that at the surface level the forms are morphosyntactically fully specified, while at the lexical level the morphological tags are suppressed and only the POS information is available.

The analyzer has parallel implementations in xfst (cf. Beesley and Karttunen 2003) and foma (cf. Hulden 2009a and 2009b).

An example derivation and a detailed excerpt from the analyzer are provided in the Appendix.

4 The Paradigms of Adjective Agreement in Standard German

German adjectives are inflected for 3 genders (masculine, feminine, and neuter), 2 numbers (singular and plural) and 4 cases (nominative, accusative, dative, and genitive). There are no gender differences in the plural.

There are three adjective agreement paradigms in Standard German: a) the strong declension (SD); b) the weak declension (WD); c) the mixed declension (MD). Additionally, there is a single uninflected[1] form that is used predicatively and is chosen as the lemma.

Below are the positive strong inflected forms of *schnell* ‘fast’:

	Masc	Neut	Femn	Plur
Nom	*schneller*	*schnelles*	*schnelle*	*schnelle*
Acc	*schnellen*	*schnelles*	*schnelle*	*schnelle*
Dat	*schnellem*	*schnellem*	*schneller*	*schnellen*
Gen	*schnellen*	*schnellen*	*schneller*	*schneller*

[1] This form is uninflected for gender/number/case but can be inflected for degree of comparison if the adjective is qualitative.

Five inflected forms (*schneller, schnelles, schnelle, schnellen, schnellem,*) are associated with 5 syncretic sets of fully specified morpho-syntactic feature values of the positive degree of a German adjective. These sets are disjunctive and their union represents the 48 possible feature values of the positive degree of an adjective in German. The same applies to the comparative and superlative degree. To make things even more complicated, the uninflected comparative form (e.g., *schneller*) is identical with some of the inflected forms for the positive degree.

5 Fine-Tuning of the Prototypical Analyzer for Different Uses

Now let us consider what can be done so that the analyzer performs optimally in all the cases of intended uses that were listed in section 1.

5.1 Intended Use in a Pipeline with a POS Tagger

The code excerpt from the lexicon in section 3 illustrates how the use of underspecification at the lexical level can reduce ambiguities in the analysis when only lemmata and POS tags are needed by the next application in the pipeline. Thus the output for *schneller* will be:

```
schnell   +A
```

and not:

```
schnell +A+Cmp+Uninfl
schnell +A+Pos+SD+Masc+Sg+Nom
schnell +A+Pos+SD+Femn+Sg+Dat
schnell +A+Pos+SD+Femn+Sg+Gen
schnell +A+Pos+SD+Pl
schnell +A+Pos+MD+Masc+Sg+Nom
```

5.2 Intended Use in a Pipeline with a Deep-Syntactic or Semantic Parser, or for Generation

On the other hand, deep-syntactic and semantic parsers will benefit from the ambiguous output listed in the previous subsection. To achieve this we need to modify the lexical level accordingly:

```
LEXICON Adj
<"+A" "+Uninfl"> # ;
<"+A" "+Pos">  AStrong;
<"+A" "+Pos">  AWeakMixed;

LEXICON AdjCmpSpl
<"+A" "+Cmp" "+Uninfl"> #;
<"+A" "+Cmp">  AStrong;
<"+A" "+Spl">  AStrong;
<"+A" "+Cmp">  AWeakMixed;
<"+A" "+Spl">  AWeakMixed;

LEXICON AStrong
…
```

Now the lexicon and the surface level are identical. The rest – inflection for gender/number/case – is modified in the same way but will not be presented due to space limitations.

This version of the lexicon can also be used for generation.

5.3 Intended Use: Modeling of Syncretism.

In this case it is not essential if the lexical level is underspecified or fully specified. It is important for the surface level to be fully specified.

The modeling of syncretism is performed in the xfst/foma file that contains the phonological and orthographical alternations and realizational rules, and possibly also rules of referral.

As we have seen in section 2, there are different types of syncretism, e.g., phonologically, lexically or morphologically determined syncretism, and different rules, e.g., symmetric or directional.

An example of phonologically determined syncretism is the collapse of the full forms of the personal pronouns for accusative and dative in the 2nd person singular in Bulgarian. The reason for this is the merger of the reflexes of the *jat*-sound (the ending for dative) and the *e*-sound (the ending for accusative). Thus *tebě* ‘you-2SG.DAT’ and *tebe* ‘you-2SG.ACC‘ collapsed into *tebe*. In this case a realizational rule is more appropriate than the use of a rule of referral:

```
+Acc|+Dat -> e ||
        +PronP +2P +Sg _ ;
```

On the other hand, the syncretism involving the forms for genitive, dative, and locative singular of 3rd-declension-class (D3) Russian nouns, e.g., *kosti* from *kost’* ‘bone’ (cf. Baerman, Brown and Corbett 2005, p. 208) is better modeled using a cascade of rules of referral, followed by a realizational rule, since this is a directional syncretism. The syncretism of dative and locative singular is well established throughout the Russian nominal declension, with locative providing the form. In this case, however, genitive provides the form for all three feature values:

```
+Dat -> +Loc ||
        +N +D2|+D3 +Sg _ ;
```

```
+Loc -> +Gen ||
        +N +D3 +Sg _ ;

+Gen -> i ||
        +N +D3 +Sg _ ;
```

For more examples of directional rules of referral, cf. Kilbury (2011) among others.

6 Conclusions

In this paper I have shown that it is possible to create a morphological analyzer that can be tailored to various intended uses with minimal effort. The most important properties of such an analyzer are: a) the surface level in the lexicon consists of tags that represent the (language specific) values of fully specified morphosyntactic features; b) the realizations are described outside the lexicon.

The fine-tuning is achieved by modifying the lexical level to the desired degree of (under)specificity and by restructuring the realizational rules, and possibly by adding rules of referral.

Acknowledgments

I am grateful to Nicolas Kimm, Natalia Mamerow, and particularly to James Kilbury for our discussions on different approaches to language modeling.

While working on this paper, I received funding from the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) in CRC 991 (SFB 991) "The Structure of Representations in Language, Cognition, and Science".

References

Matthew Baerman, Dunstan Brown, and Greville G. Corbett. 2005. *The Syntax-Morphology Interface. A Study of Syncretism.* Cambridge University Press, Cambridge, UK.

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology.* Palo Alto, CA: CSLI Publications.

Mans Hulden. 2009a. *Finite-State Machine Construction Methods and Algorithms for Phonology and Morphology.* PhD Thesis, University of Arizona.

Mans Hulden. 2009b. Foma: a Finite-State Compiler and Library. In: *Proceedings of the EACL 2009 Demonstrations Session*, 29-32.

James Kilbury, 2011. *Implementing Comparative Reconstruction with Finite-State Methods*. Paper presented at the University of Düsseldorf. (Accessed on 10.04.2012, at http://user.phil-fak.uni-duesseldorf.de/~kilbury/publ.htm)

MULTEXT-East morphosyntactic specifications, Version 4, 2010. *MULTEXT-East Morphosyntactic Specifications, Version 4, 2. Common MULTEXT Specifications*. (Accessed on 10.04.2012, at http://nl.ijs.si/ME/V4/msd/html/msd.common.html

Gregory Stump. 2001. *Inflectional morphology.* Cambridge: Cambridge University Press.

Heike Telljohann, Erhard W. Hinrichs, Sandra Kübler, Heike Zinsmeister, and Kathrin Beck. 2009. *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z)*

Arnold Zwicky. 1985. How to describe inflection. In *Proceedings of the eleventh annual meeting of the Berkeley Linguistics Society*, 372-386

Appendix. Derivations of *fettarm* 'low-fat'.

Below are the upper and the lower side of the adjective *fettarm* that has been through several continuation lexicons of the lexc-lexicon. The desired analysis output is on the upper side, while the morphotactics is on the lower side. The realizational rules operate only on the lower side.

Upper:	fett#arm	0	+A	0	0	0	0	0
Lower:	fett1rm	+Uml	+A	+Pos	+SD	+Msc	+Sg	+Nom

Rule (1) below defines the realization of *1* as *ä*:

```
(1) define UML  [1 -> ä || _ $["+Uml" "+A" ["+Cmp"|"+Spl"] ] ;
```

Since the conditions are not met, the lower-side string remains unchanged. The next rule (2) defines the realization of the adjectival suffix *-er*:

```
(2) define AEr   [ [ [ "+SD" | "+MD" ] "+Masc" "+Sg" "+Nom" |
                     "+SD" "+Femn" "+Sg" [ "+Dat" | "+Gen" ] |
                "+SD" ["+Masc" | "+Femn" | "+Neut"] "+Pl" "+Gen" ]
                     -> %+ e r || _ .#.] ;
```

The string is now: fett1rm+Uml+A+Pos+er. Rule (3) defines the realization of the comparative *–er* suffix:

```
(3) define ACmp    [ "+Cmp" -> %+ e r || _ [%+|"+Uninfl"] ] ;
```

Since the conditions are not met, the surface string remains unchanged. The next rule (4) defines the realization of *1* as *a*:

```
(4) define UMLT      [1 -> a, 2 -> o, 3 -> u, 4 -> A] ;
```

The lower-side string is: fettarm+Uml+A+Pos+er. The last rule (5) deletes the + and the remaining tags that were not used in the realization:

```
(5) define TagDel     [RestTag -> 0 ] ;
```

The lower-side string is now: fettarmer.

A lower-side string fett1rm+Uml+A+Cmp+Uninfl will render fettärm+Uml+A+Cmp+Uninfl after the application of rule (1), fettärm+Uml+A+er+Uninfl after rule (3), and fettärmer after rule (5).