LEMPAS: A Make-Do Lemmatizer for the Swedish PAROLE-Corpus

Silvie Cinková and Jan Pomikálek

Abstract

LEMPAS, the lemmatizer for the Swedish corpus PAROLE, came into existence as a by-product of running the Sketch Engine (Kilgarrif et al., 2004) on Swedish, since many of the desirable features of the Sketch Engine, such as building word sketches, are only available for lemmatized corpora. We did not have access to any Swedish lexical sources and the time allowed for the lemmatization was very limited. Consequently, the lemmatizer had no great design ambitions. Initially, we were only attempting to bring related forms together under a pre-lemma, using general rules, and avoiding explicit lists where possible. When the initial rules gave surprisingly good lemmatizations of nouns, verbs and adjectives, we decided to transform the pre-lemmas into real lemmas. The improved lemmatizer made a very good impression. We have tested the program on the manually lemmatized Stockholm-Umeå Corpus (SUC), and have analyzed the results.

1 Introduction

This paper presents LEMPAS, a rule-based lemmatizer for the Swedish PAROLE corpus. We built it without being aware that the SUC-based statistical tagger/lemmatizer (Hajič, 2002) was already in existence; and its 94.72 % accuracy would have been more than adequate for our purposes. Nor had we succeeded in obtaining any Swedish-made lemmatizer. On the other hand, we believe that a properly designed rule-based lemmatizer can achieve higher accuracy than a data-based one. In our work we have taken the first steps towards achieving such a program.

When building LEMPAS, we did not have access to any Swedish machine-readable lexicon, and the time for this task was quite limited. Initially, the lemmatizer only had to be good enough for computing collocational relations in PAROLE with the tool *Sketch Engine* (Kilgarrif et al., 2004). Actually, for running the Sketch Engine, the lemmatizer only had to be able to bring together related inflectional forms of nouns and verbs under one common form, which in turn did not even need to correspond to the headword under which the lexemes are codified.

The initial rules returned surprisingly good results, and a few random corpus queries revealed further subtypes of inflection for which we introduced several complementary rules. The result (i.e. bringing together related inflection forms) exceeded our expectations, and we decided to make further improvements. This mainly involved making our "pre-lemmas" comply with conventional headwords.

The improved lemmatizer made a very good impression. To examine it more closely, we have performed a quantitative comparison of the LEMPAS-lemmatized PAROLE with the manually lemmatized Stockholm-Umeå Corpus (SUC), and analyzed the results.

2 The Corpora Used

2.1 PAROLE

The Swedish PAROLE is a corpus of modern Swedish texts that comprises more than 19 million running words. PAROLE belongs to Språkbanken, the set of corpora at Språkdata, University in Gothenburg,

Sweden, and is available at `http://spraakbanken.gu.se/`. The PAROLE corpus was built within the EU project PAROLE (completed in 1997), which aimed at creating a European network of language resources (corpora and lexicons).

2.1.1 The tagset of PAROLE.

The tagset identifies the parts of speech. It considers the following categories for nouns, verbs and adjectives respectively:

Nouns

- common (NC) vs. proper (NP)
- gender: neuter (N) vs. uter[1] (U)
- number: singular (S) vs. plural (P)
- case: "nominative" (N) vs. genitive
- definiteness: definite (DS) vs. indefinite[2] (IS)

Verbs

- verb vs. adjective[3]
- tense/mode: present (II) vs. preterite (IP) vs. supine[4] (IU) vs. imperative (M0) vs. infinitive (N0) vs. present and past conjunctive (SI, SP)
- voice: active (AS) vs. deponential (SS)[5]

Adjectives

- participle (AP, AF) vs. adjective (AQ)
- degree: positive (P) vs. comparative (C) vs. superlative (S)
- gender: masculine (M)[6] vs. uter (U) vs. neuter (N)
- number: singular (S) vs. plural (P)
- case: "nominative" (N) vs. genitive (G)
- definiteness[7]: definite (DS) vs. indefinite (IS)

The irrelevance of a category is indicated by 0 at the appropriate position.

[1] merged masculine and feminine

[2] Swedish employs the postpositive definite article, which is attached as a suffix to the noun.

[3] Past (AF) as well as present participles (AP) are regarded as adjectives.

[4] The supine form (supinum) of the verb in Swedish is used with the auxiliary verb *ha* ('to have') to form the perfective.

[5] Deponential verbs have the suffix *–s*. Few verbs exist only as deponentials but generally, the deponential form indicates either reciprocity or passive voice.

[6] The ancient masculine form is not very common but still occurs in certain usages.

[7] Only indicated in singular. The definiteness value of the adjectival attribute must agree with the definiteness value of the governing noun.

2.2 SUC

The SUC corpus comprises approximately one million running words. It was both tagged and lemmatized manually. Its tagset (Ejerhed et al., 1992) differs from that of PAROLE but either of them can be easily transferred into the other. The comparison of the respective tagsets is available at the PAROLE website.

3 Building LEMPAS

The linguistic task to be processed by the Sketch Engine only required the lemmatization of nouns and finite verbs. Besides, we added a partial lemmatization of adjectives; i.e. we lemmatized only tokens with the following tags: NC.* (common nouns), A.* (adjectives and participles) and V.* except V@S.* and V@M.* (verbs except imperatives and conjunctives). We also systematically ignored numbers (M.*), proper nouns (NP.*), pronouns (P.*) and adverbs (R.*).

LEMPAS comprises a sed script and a complementary Perl script. The sed script gathered related inflection forms, while the Perl script corrected the pre-lemmas to comply with headwords. The lemmatization is tag-dependent. We used a simple regular expressions syntax to build the basic lemmatization rules. This made the implementation straightforward enough to be performed by linguists with limited computer skills. We are well aware that the implementation could be much more efficient if appropriate searching data structures were used for finding lemmatization rules by word endings.

3.1 The Sed Script

The data of the PAROLE corpus to be processed with the Sketch Engine had the following structure: one token per line, followed by a tag separated by a tab. A lemma string was to be inserted inbetween. For example, the token *katterna* ('cats-the') would have looked like *katterna NCUPN@DS* and would be replaced by *katterna katt NCUPN@DS*. The "find-replace" sed-structure hosts linguistic rules that decompose the "word" string into segments to be preserved, omitted or modified in the "lemma" string, which is newly created by the replacement. The next sections describe the linguistic rules in more detail.

3.2 Rules for Nouns

The script groups the rules approximately according to declension types as listed by Nylund (Nylund and Holm, 1993). Several declensions go across genders; therefore genders can only be read from tags.

Several declensions contain an additional rule that affects the indefinite singular. This rule chops off the last character from stems ending with *–e*. As the rules for definite singulars and both plurals are unable to determine whether the lemma is supposed to end with *–e* or not, we decided to handle the lemmas as if Swedish had no words ending with *–e*. Thanks to this rule, the word *stavelse* gets the lemma *stavels* in all forms. Table 1 lists some examples of the basic rules for nouns.

The rules for genitives are identical with the rules for nominatives except for the *–s* added to the respective endings and G replacing N at the case position. We neglected words ending with *–s*, *–z*, and *–x*, which do not attach *–s* in the genitive. We assume that genitives of such words would mainly occur in proper nouns, which lemmatization ignores in general.

On the other hand, we have paid attention to types of common nouns that do not fit the basic rules.

One of the important rule restrictions has been applied to nouns whose plural forms and definite singular form in neuters follow a stem *l*, *r* or *n*, which is neither duplicate (*stället – ställe*) nor preceded by a vowel (*signaler – signal*): *muskler*, *mörkret*. These nouns have usually dropped their *e* in the indefinite singular: *muskel*, *mörker*. We have thereby ignored words that originally had no *e* in their stem. So far we have found and listed the counterexamples *moln*, *karl*, *kärl*, *sorl*, *porl*, *regn*, *ugn*, *agn*, *vagn*, *lögn*, *stygn*, *lugn* and *dygn*. We also have listed the noun *morgon* (plural *morgnar*).

Table 1: Examples of lemmatization rules for nouns

ending	tag	rule	example
–an	NCUSN@DS	–n deletion	flickan → flicka
–or	NCUPN@IS	–or deletion	flickor → flicka
–ar, –er	NCUPN@IS	–ar, –er deletion	katter → katt, stolar → stol
–en	NCUSN@DS	–en deletion	katten → katt, stolen → stol
–[iouyöää]n	NCNPN@IS	–n deletion	hjärtan → hjärta, möten → möte
–[iouyöää]t	NCNSN@DS	–t deletion	hjärtat → hjärta, mötet → möte
–n	NCNPN@DS	–n deletion	husen → huse
–et	NCNSN@DS	–et deletion	huset → hus
–[iouyöää]n	NSUSN@DS	–n deletion	byrån → byrå

Table 2: Examples of lemmatization rules for verbs

ending	tag	rule	example
–ar	V@IPAS	–r deletion	klarar → klara
–er	V@IPAS	–r deletion, –a insertion	läser → läsa
–[öäiouåy]r	V@IPAS	–r deletion	syr → sy, tror → tro

We add an *m* to lemmas with the singular definite ending *–mlen*, *–mlet*: *himlen – himmel*, *skramlet – skrammel*. We also add an *m* to lemmas with the *–ar* plural endings following an *m* that is neither duplicate (*dammar – damm*) nor preceded by a vowel (*kramar – kram*): *kamrar → kammare* (*somrar → sommar* is listed). This rule ignores nouns ending with *–mer* in the singular indefinite as the plural forms of these nouns (*glimmer*, *flimmer*, *bekymmer*) only occurred as tagging errors. Another m-rule chops off *m* in neuters in which the definite endings follow a double *m* (*hemmet → hem*, *programmet → program*). Only suggesting the *l*, *r*, *m*, *n* subtypes, this paper does not list rules for all inflection forms, though they are present in the script.

Another problematic group of nouns are loan neuters with *–er* ending in plural, which can have two different suffixes in singular: *–ium* vs. *–eri*, e.g. *podier → podium* vs. *skafferier → skafferi*. Our rules were able to correctly lemmatize nouns ending with *–orier*, *–arier* and *–eer*. The types *mysterier → mysterium* and *podier → podium* could not be distinguished by rules and were therefore processed by the Perl script. There was a similar problem with plural uters ending with *-ier* (*serier → serie* vs. *harmonier → harmoni* vs. *historier → historia* vs. *irakier → irakier*). To a certain extent, these types have also been resolved by the Perl script.

3.3 Rules for Verbs

Irregular, modal and auxiliary verb forms have been listed. Conjunctives and imperatives have been ignored by lists as well as by rules. Examples of rules for present active forms of regular verbs are given in Tab. 2.

The rules for deponential forms are almost identical with the rules for active forms, just an *–s* suffix follows the conjugated form.

Table 3: Rules for building dictionaries

dict #	*ending*	*tag*	*delete-ending*	**example**
1	–e	NC.SN@IS	–e	möte, add möt
2	–ia	NCUSN@IS	–a	historia, add histor
3	–um	NCNSN@IS	–um	podium, add podi
4	–ier	NCUSN@IS	–er	irakier, add irak
5	–are	NCNSN@IS	–are	läkare, add läk

3.4 Rules for Adjectives

The tagsets of both PAROLE and SUC regard participles as a subset of adjectives. The present participle forms differ only in case (nominative vs. genitive) and the nominative is the lemma (e.g. *asylsökandes – asylsökande*, *oberoendes – oberoende*). Nominative forms were copied into the lemma and *–s* was deleted from the genitives.

The lemma of a perfect participle is identical with the indefinite singular uter nominative (e.g. *avklarade – avklarad*) as is found in adjectives. Perfect participles of irregular verbs acquire different endings than those of regular verbs. The genitive is marked by the *–s* suffix. Rules are set for all genders (incl. masculine), both numbers, and for both definite and indefinite forms. They cover the following types: *klarad*, *berörd*, *köpt*, *ansedd*, *bruten*, *välkommen*.

Rules for adjectives include gradation, itemizing the commonest irregular forms, and seeking to cover systematic stem changes like the noun rules.

3.5 The Perl Script

As mentioned above, some of the rules return only pre-lemmas rather than real lemmas. To amend this, we used a simple Perl script for postprocessing. The script operates in two steps. In the first step, the whole corpus is read and a set of dictionaries is built of words meeting certain conditions. Then the corpus is gone through again and some of the lemmas are modified according to the dictionaries.

The following strategy is used when building the dictionaries:

if word ends with *ending* **and** its tag equals to *tag* **then**
add the word to the dictionary with the *delete-ending* deleted

The values of the parameters for each of the dictionaries are listed in the Tab. 3.The rules for modifying lemmas according to the dictionaries include:

if pre-lemma in $dict_1$ **and** tag matches *NCU[SP][NG]@[ID]S* **then**
lemma := pre-lemma + *–a*

if pre-lemma in $dict_2$ **and** tag matches *NCN[SP][NG]@[ID]S* **then**
lemma := pre-lemma + *–um*

if word ends with *–arna* **and** tag=*NCUPN@DS* **and** word base in $dict_3$ **then**
lemma := word base + *–are*

if word ends with *–arnas* **and** tag=*NCUPG@DS* **and** word base in $dict_3$ **then**
lemma := word base + *–are*

if pre-lemma in $dict_4$ **and** tag matches *NCU[SP][NG]@[ID]S* **then**
lemma := pre-lemma + *–er*

Table 4: Lemmatization results evaluation

	all words			unique words		
	correct	errors	accuracy	correct	errors	accuracy
common nouns	221201	13600	94.21 %	56987	5578	91.08 %
finite verbs	113452	22006	83.75 %	7707	1123	87.28 %
adjectives	56066	16670	77.08 %	10742	1426	88.28 %
present participles	5414	8	99.85 %	1462	7	99.52 %
perfect participles	11394	1096	91.22 %	4451	429	91.21 %
all	407527	53380	**88.42 %**	81349	8563	**90.48 %**

3.5.1 Example.

The *e*-eliminating sed rule lemmatizes the singular indefinite nominative noun *möte* as *möt*, which unifies it with the inflected forms *mötet*, *mötes*, *mötets*, *möten*, *mötens*, *mötena*, *mötenas*, which all have been lemmatized as *möt* by the basic rules. The Perl script detects the difference between the word string and the lemma string in the indefinite nominative singular (*möte NCNSN@IS möt*) and includes *möt* in the dictionary as $case_1$. Then it searches all of the tokens for this string in the lemma and groups the relevant ones under the $case_1$-parameter. Finally, the lemmas of all inflection forms of the word *möte* get *–e* attached, i.e. the lemma is corrected in all inflection forms.

The Perl script can neither resolve ambiguous forms nor orthographic variants. It simply selects the more frequent indefinite singular form to be the lemma. For example, the noun *herr* ('Mr.') will be lemmatized as *herre* due to the predominant form *herre* ('sir', 'Lord'). Initially, one single occurrence of the given word string ending with *–e* was enough to include the word in the dictionary. However, rare archaic and colloquial word forms turned out to harm the lemmatization. For example, the extremely common noun *hus* ('house') was lemmatized incorrectly (with *–e*) due to its archaic inflection form *huse*, nowadays only used in the phrase *man ur huse* ('altogether')! To avoid this nuisance, the script was enhanced with a frequency comparison of *e*-ending vs. non-*e*-ending indefinite singular nominative word strings.

Besides that, forms lacking their singular indefinite counterpart in the corpus are never lemmatized correctly. This is often the case of occasional compounds as well as group names such as *irakier* and *indier*, which typically occur in the plural.

4 Results Evaluation

The manually lemmatized SUC was used for evaluating our lemmatizer. As already mentioned, we focused on lemmatization of some parts of speech only. When evaluating the results, we ignored any other word types. Out of the 115,228 SUC unique words 89,912 were analyzed, i.e. 78 %.

LEMPAS was run on SUC and its results were compared to the original lemmatization. In addition to overall results we report the number of the correctly and incorrectly lemmatized words for each of the following word groups: common nouns (NC.*), finite verbs (V@I.*), adjectives (AQ.*), present participles (AP.*) and perfect participles (AF.*). In Swedish, the lemma is always uniquely determined by the word and its POS-tag. Consequently, if a word appears in the corpus with the same POS-tag repeatedly it is always assigned the same lemma. Therefore we report the results both for all the words in the corpus, and for unique words only. In the latter case the evaluation is done as if every word appeared in the corpus with the same POS-tag only once.

For several reasons we did not attempt to compare the LEMPAS to Hajič's statistical tagger/lemmatizer. First, the statistical tagger was trained on the SUC corpus, which greatly favours it if the comparison

was to be done on the SUC corpus. Unfortunately, we did not have any other manually tagged and lemmatized corpus available. Second, the statistical tagger performs a different task then LEMPAS. While LEMPAS only lemmatizes an already tagged corpus, Hajič's tagger tags the corpus first and only then lemmatizes it. Therefore LEMPAS would have the advantage of correct tagging, while its competitor would have to guess it. This makes the two tools hard to compare.

5 Conclusion

LEMPAS came into existence as a make-do, thus without great ambitions. However, we dare to offer it to the public as it is, at least as an approved supporting tool for others who happen to use the Sketch Engine on Swedish data. We also believe that it has the potential for worthwhile development.

6 Acknowledgements

This work has been supported by LC536, GAUK 489/2004, IS-1ET101120413, and IS-1ET201120505, and it was only enabled by the Språkbanken team's generous permission to download the entire corpus.

References

Ejerhed, E., G. Källgren, O. Wennstedt, and M. Åström. 1992. The linguistic annotation system of the stockholm–umeå corpus project. Technical Report 33, Dept. of General Linguistics, University of Umeå, Umeå, Sweden.

Hajič, Jan. 2002. Se030107x - a statistical tagger/lemmatizer based on the suc corpus. Personal communication 2006-03-19.

Kilgarrif, Adam, Pavel Rychlý, Pavel Smrž, and David Tugwell. 2004. The sketch engine. In *Proceedings of the Eleventh EURALEX International Congress*, pages 105–116, Lorient, France.

Nylund, E. and B. Holm. 1993. Deskriptiv svensk grammatik.