Keeping the Speller Lexicon up-do-date for an Inflective Language

Heiki-Jaan Kaalep
University of Tartu
Vaba 19, Tartu 50114, Estonia
Heiki-Jaan.Kaalep@ut.ee

Abstract

Updating a lexicon is a laborious and time-consuming task. In the context of keeping one's proofing tools up to date, however, this task seems to be impossible to avoid. The paper concentrates on updating a speller lexicon for an inflective language, using Estonian as an example. The Estonian speller was first licensed by Microsoft in 1995, and the lexicon has gone through several updates since then, growing from 31,000 to 56,000 words. Based on the notion of stable and unstable inflectional classes, it has been possible to automate the updating process to a great extent. However, a normal text always contains words that have not been met in previous texts earlier, and are thus missing from the speller lexicon also, no matter how hard we have tried to improve it. This is a fundamental feature of the human language. A practical solution would be to add a feature to the user dictionary that would make it possible to guess the inflectional class of the added word, generate all the inflectional forms of the word, and add them to the user dictionary. This would be clearly preferable to the current limitation that the inflected forms can be entered by the user only one by one.

Introduction

Updating a lexicon is a laborious and time-consuming task. In the context of keeping one's proofing tools up to date, however, this task seems to be impossible to avoid. For an inflected language, this task is made more difficult by the need to include morphological information in the dictionary, in addition to the headwords, to guarantee the possibility to analyze inflectional forms. The speller for Estonian, an inflected language, was first licensed by Microsoft in 1995, and the lexicon has gone through several updates since then. Looking back, it is obvious that keeping a lexicon up-to-date is not simply a question of what should be added to the lexicon and what deleted. One must automate the updating process, especially for an inflected language, where every word belongs to some inflectional class and consequently must have some marker attached to it. One also cannot help wondering if there is any hope that some day the lexicon will be perfect and cover all the correct words of a text. Otherwise, what could be done to reconcile the irritated user after (s)he has encountered yet another normal word form, flagged by the speller?

Why revise a lexicon?

There are at least two reasons why one has to revise the lexicon.

First, every time we add a new component to the proofing tools, this component will require an update of the speller's lexicon. For example, think about a thesaurus or a bilingual dictionary. It is common knowledge that dictionaries differ in their choice of headwords. It is not only that some dictionaries are bigger; interestingly, it is a norm that a smaller dictionary contains some words that are missing from the bigger one. If dictionaries are meant to meet different needs, as happens in the case of a bilingual dictionary and a spelling dictionary, their lists of headwords are bound to differ considerably. It would be weird if the thesaurus or a bilingual dictionary suggests a word that the speller does not recognize and will thus flag as a mistake.

Second, languages change and develop over time. New words are adopted constantly, and they should be added to the lexicons.

Regularities in inflectional classes

When adding a new word to the lexicon of an inflective language, we must define its inflectional type. The inflectional type defines the set of possible inflectional affixes and the pattern of morpheme alternations for the word. Fortunately, the morphological system of a language is more regular than it looks at first glance. To see this, we must have a look at language change. The following citations from Wolfgang U. Wurzel who represents a research paradigm called Natural Morphology [4], and from an academic grammar of Estonian [2], serve to give the background. The terminology they use conveys an idea about a certain remarkable aspect of applicability of morphological rules, although different authors and traditions use different terms, e.g. "active and passive morphology" [2]; "dynamic and static morphology" (W. U. Dressler); "stable and unstable inflectional classes" [4].

Wolfgang U. Wurzel says:

"It can be said that the word as a lexical unit is defined by the mutual assignment of a semantic and phonological structure; the two sides in their interaction constitute the word. In contrast, the inflectional-morphological properties of a word, i.e. its membership in an inflectional class, are not constitutive; rather, they act as some kind of accompanying conditions. In many inflected languages, particularly in all strictly agglutinative languages, there are no different inflectional classes at all, cf. […] Turkish noun declension. Actually, (inflectional-) morphological properties are nothing but operational instructions for using words to form sentences. In principle, they have to be learned in addition to the meaning and sound form of words, so they require an additional learning expenditure which, strictly speaking, is unnecessary for the functioning of the language. This expenditure can be kept relatively low because morphological properties are dependent on the independently existing extramorphological properties of words. Therefore morphological properties tend to depend on phonological and/or semantic-syntactic properties. Phonological properties which can be utilized accordingly include the 'ending' of the basic form or the vowel of the basic morpheme; semantic-syntactic properties include gender or features like 'person', 'animateness' and 'kinship' in the noun and 'modality', 'transitivity/intransitivity' and 'punctuality' in the verb etc.

[…]

If e.g. in Russian a noun ends in /*a*/, then it has /*i*/ in the G. sg., /*e*/ in the D. Sg. etc., cf. N. *sobaka* 'dog' – G. *sobaki* – D. *sobake* etc.

[…]

For most inflectional languages, extramorphological properties and inflectional class typically do not coincide; frequently, two or more inflectional classes contain words with the same extramorphological properties. Such inflectional classes will be called complementary classes. A good example involves German nouns with a phonologically short, phonetically medium-length word-final vowel (except /*e*/), where there are four complementary classes side by side, cf. e.g. *Kino* 'cinema' – Pl. *Kino-s*, *Fresko* 'fresco' – Pl. *Fresk-en*, *Cello* 'violoncello' – Pl. *Cell-i* and *Schema* 'schema' – Pl. *Schema-ta*. It should be pointed out that each variant of plural formation has a different status for German speakers, as clearly demonstrated by language change, treatment of neologisms, child language etc. For a long time, words have been changing from the class of *n*-plurals to that of *s*-plurals, cf. *Arom-en* 'aromas' > *Aroma-s* […]. There are no converse changes of the type *Kino-s* 'cinemas' > **Kin-en*. All neologisms (including loan-words) of the common language join the class of*s*-plurals, cf. *Disko-s*, *Pizza-s* and *Ufo-s*. […] Plurals formed with /*i*/ and /*ta*/ are today only optional variants beside *s*- and *n*-plurals.

[…]

It has become customary to distinguish between productive and unproductive inflectional classes in inflectional morphology. The main criterion for the productivity of inflectional classes is considered to be their 'openness': They are 'open' to new members, while the unproductive classes remain 'closed' in this sense. […] Such classes are 'open' […] for words with specific extramorphological properties; they are not 'open to all sides'. For instance, the 'openness' of the *s*-plural class of nouns of the type Kino 'cinema' is beyond question […]. But this 'openness' does not apply to nouns with any extramorphological properties, e.g. not to nouns in /e/, cf. *(der) Taiwanese* 'Taiwanese' – Pl. *(die) Taiwanese-n* (*(*die Taiwanese-s*))." [4]

An academic grammar of Estonian states:

"If a language has several different ways to decline a noun or conjugate a verb, then the base form of the word itself must contain information for selecting the right set of inflectional rules.

In Estonian, this type of information is dependent on the phonological-derivative features of the base form of a word. […] For instance, if a noun is a monosyllabic word ending with a consonant, its G. pl. is formed by adding /i/ to the base form and /*de*/ for marking plural, e.g. *kass* 'cat' – G. pl. *kass-i-de*; […] if a noun has been derived by adding an affix *-kas*, its G. pl. is formed by deleting the final /*s*/ from the base form and adding /*te*/ for marking plural, e.g. *pastakas* 'ballpoint pen' – G. pl. *pastaka-te*. […]

Alas, Estonian contains many words with a similar phonological-derivative structure that nevertheless behave morphologically differently. For instance, *sari* 'sareei, an Indian garment' – G. sg. *sari*; *sari* 'series' – G. sg. *sarja*. […]

Because of the latter, it would be sensible to divide Estonian morphology in two: active and passive morphology.

Active morphology covers most of the lexicon; active morphological rules are triggered by the phonological-derivative features of the base form, without any need for additional information about the word.

Passive morphology covers instances where the base form does not define uniquely, which rules should be applied for forming the inflected forms. The rules of passive morphology have their roots in the history of the language, e.g. phonological change. From today's viewpoint, the rules of passive morphology may be regarded as lexicalised. This means that a speaker must know beforehand the rule for generating an inflected form of a word, and cannot deduce it automatically from the structure of the base form. […]

Usually, an active morphological rule is also productive, i.e. it is used for inflecting new words (loanwords, derived words, neologisms) and tends to be used in colloquial Estonian for words which are subject to passive rules in normative literary Estonian." [2]

A word, inflected according to rules of passive (static) morphology, i.e. a word belonging to an unstable inflectional class, must be frequent enough in texts, to keep its inflectional pattern. Infrequent words cannot belong to unstable inflectional classes (except for a short time, obviously). The necessity of being frequent means that all the words belonging to unstable inflectional classes are likely to get included in dictionaries of the language.

If we compare the number of different words belonging to stable and unstable inflectional classes, it is evident that unstable classes are very small, compared with stable ones. This property can be used to determine the nature of an inflectional class, without the need to examine the change of language in time.

All the above applies to simplex words, of course. It may be that a word belonging to an unstable inflectional class participates in productively coined compound words very actively, and if we don't look at the inner structure of words, we get an impression that an unstable inflectional class is productive. For instance, Estonian *mees* 'man' belongs to an inflectional class of 29 simplex-word members only, but the number of compound words with *mees* in the speller lexicon is over 200, e.g. *kaup+mees* 'trades+man', i.e. 'merchant'; *kala+mees* 'fisherman'; *iga+mees* 'everyone' etc.

Estonian, an example of an inflective language

What W. U. Wurzel has said about German and Russian inflectional classes, can also be applied to Estonian quite successfully, although Estonian belongs to a different language family.

Estonian is a Finno-Ugric language, spoken by about one million people. It is an inflective language: a declinable word (noun, adjective, numeral or pronoun) has typically 28 or 40 different inflectional forms (depending on the inflectional type of the word); a verb has typically 47 different inflectional forms. Estonian inflection involves appending inflectional affixes to a stem, as well as alternations in the stem itself. A word has often more than one stem variant, e.g. *padi* 'pillow', *padja-s* 'in a pillow', *patja-des* 'in pillows'. To make things more complicated, new Estonian words can be formed freely and productively by derivation and compounding. Derivation is a process where adding an affix (a suffix, or less frequently, a prefix) produces a new morphological word having its own inflectional paradigm. The formation of Estonian compounds is quite free: inflected words, stems, truncated stems or derived words belonging to any word class (excluding conjunctions and acronyms) may be glued together to form new compound words, although not all combinations are allowed. Up to 5 stems may be glued together, e.g. *raud+tee+üle+sõidu+koht* 'railway crossing'. About 8% of the word forms in a running Estonian text are derived words, and more than 12% are compound words. In newspaper and scientific texts the figures are even higher.

So a speller of Estonian cannot consist solely of a lexicon and a look-up module. It has to contain an algorithm for combining elements – stems, prefixes, derivational affixes (suffixes) and inflectional affixes (endings) – for forming correct word-forms. Thus the speller obtains the ability to recognize regularly formed new words, in addition to words that exist in the lexicon already.

Adding new words to the lexicon

The lexicon of the Estonian speller was initially based on the Morphological Dictionary of Estonian by Ülle Viks [3]. Initially it contained 35,000 simplex words, but after having dialect, archaic or regularly derived words removed, it shrunk to 31,000.

Over 8 years, the speller's lexicon has grown by 25,000 words, from 31,000 to 56,000. Most of the new words, 19,000, came from the thesaurus which was included in the Estonian proofing tools for MS Office in 1997. Of these, 12,000 are compound words and 3000 are expressions. Processing various text corpora, and feedback from users and Microsoft, have resulted in adding 6,000 words, including 4,000 compounds.

None of the 25,000 words had any marker attached to them initially, to denote their inflectional class. Fortunately, defining the inflectional class of an Estonian word can be automated to a great extent, as exemplified by the following algorithm.

First, pick out the unknown words from the list of dictionary headwords, and pass them to a morphological guesser. The guesser is a program that guesses the lemma form, the morpheme boundaries, and the grammatical categories (like part-of-speech, number and case) of a word, based on the word's final letters and syllable structure. Although a guesser is typically created for guessing unknown words in a text, it can be used for guessing the structure and part of speech of dictionary headwords as well.

The program takes into account the final letters of the word and the number of syllables; it does not take into account the context of the word.

During the guesswork the program checks if the word could belong to one of the following categories:

1. an abbreviation (up to two letters or a 'word' consisting of consonants only; or a word consisting of upper-case letters with a possible attachment of lower-case inflectional affix);
2. a spelling error - a word that will be analyzable after the mistake is corrected (e.g. there is no space between words, or there is a sequence of three identical vowels);
3. a proper noun;
4. a derived or compound word with either a rare formation pattern, or one including a simplex word not included in the lexicon;
5. an unknown simplex word – a noun or a verb (the judgment is made on the basis of a possible inflectional affix of the word and the number of preceding letters and syllables).

In case of dictionary headwords, the guesser normally classifies the words into categories 4 and 5.

For example, we have new simplex words, which have been adopted from some other language. E.g. *euro* was 8 years ago used only in colloquial Estonian for ‘European’, and had to be kept away from the speller’s lexicon. Now it is the name of a currency, and a frequent component in well-formed compound words, and thus belongs to the lexicon. An example of a compound word that contains a previously unseen component *mail* would be *maili+uputus* ‘mail-flood’.

After this stage, we have two types of words: first, derived and compound words, the inflectional class of which is determined by the final component (derivational suffix or word stem), and second, simplex words. Derived and compound words belong to the same inflectional class as their last component, and can thus be automatically classified.

Now we are left with simplex words. Fortunately, new simplex words belong to only a limited set of inflectional classes, in accord with what W. U. Wurzel has described about German in [4] and as predicted by [2] about Estonian. When classifying a new word, it is sufficient to take into account its syllable structure and some final letters; based on these, one can automatically classify the simplex words.

Initially, the Estonian speller lexicon contained 26 different classes for 22,000 declinable words and 12 classes for 6300 verbs. In contrast, all the new simplex declinable words, added to the lexicon during the last 8 years (4600 altogether), belong to only 12 classes, and all the simplex verbs (500) belong to 4 classes. At the same time, there are only 900 simplex nouns, adjectives, numerals and pronouns, and only 200 simplex verbs in unstable inflectional classes. They were all present in the first version of the speller’s lexicon already. It is interesting to note that the number of irregular verbs in English is about 170 and in German about 210.

By following the outlined steps above, one can transform a list of headwords into a lexicon with the necessary morphological information for analyzing and synthesizing all the inflectional forms of the headwords.

A lexicon cannot be perfect

The fact that the inflectional class of a word can be predicted, based solely on the form of the word, indicates that there is a strong need for predictability like this. Natural language texts typically contain words, missing from previously compiled lexicons. Unfortunately, it is a fundamental feature of the human language, as shown by H. Baayen in [1].

“Word frequency distributions are Large Number of Rare Events (LNRE) distributions, distributions characterized by the presence of large numbers of words with very low probabilities of occurrence. In the British National Corpus, for instance, more than half of all types have a sample relative frequency of .00000001. Due to the large number of rare words, the sample size N has to be extremely large for the asymptotic properties of the distribution to emerge. In practice, almost all samples of words are located in the LNRE zone, the range of sample sizes where the vocabulary size is still increasing, and where the numbers of hapax legomena (words that occur only once in the corpus), dis legomena (words that occur twice), etc., are non-negligible.” ([1], p.54-55)

In other words, even after we have seen tens of millions of words, we keep stumbling on large numbers of previously unseen words, and the growth rate of the vocabulary is largely unpredictable. Figure 1 serves as an illustration of the vocabulary growth (V), dependent on the corpus size (N).

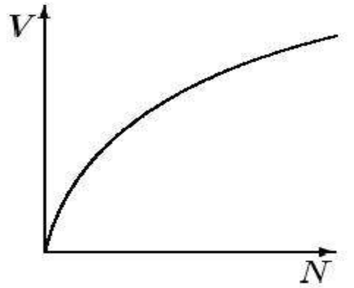

Figure 1. Dependency of the vocabulary size V on the text size N.

It is notable that the growth curve, although losing in steepness as the vocabulary grows, is still not asymptotic. This implies that the size of the vocabulary is infinite. Baayen also shows that for all values of N, the number of hapax legomena exceeds the number of dis legomena, the number of dis legomena exceeds that of the tris legomena, etc.

Hapax legomena make up roughly half of the vocabulary of a text corpus, representing only a few percents of the running words (the exact figures are dependent on the text size, text type, language, and the way we count the words). This is roughly proportional with the amount of words of the same text corpus, which are missing from a pre-compiled lexicon, and thus would be flagged as errors by a speller.

A test on a newspaper corpus of "Eesti Ekspress" (the biggest nation-wide weekly newspaper in Estonia) from 1999, collected from the Web (http://www.ekspress.ee) and containing 219,000 word tokens, revealed the following.

10,000 (4.57%) of the word form tokens, including those with non-standard orthography and typographical errors, were not recognized by the lexicon-based morphological analyzer. They fall into the following categories: about 66% of unknown tokens are proper nouns; 10% are common nouns; 9% are punctuation marks that occur in some non-standard form (e.g. dash); 8% are abbreviations; 1% are various combinations of numbers and letters; 1% are adjectives, verbs and adverbs; 5% are foreign words, web addresses, and other sequences of symbols for which it is difficult to offer any reasonable analysis.

In view of the above, it is not surprising that previously unseen simplex words can be automatically classified into inflectional classes. Texts always contain previously unseen words (think of the number of previously never met proper names in a newspaper!), and people have to be able to process them without too much effort. In case of an inflectional language, people have to be able to deduce the base form from the inflected form they meet in the text, as well as to inflect the word without errors, if they happen to talk about it. So the inflectional system of the language has to be regular for the new words, so that people could communicate seamlessly.

(The requirement of regularity does not mean, however, that it is always met in real life. A confusing matter in Estonian is the declension of proper names which may entail an option to transform the word stem according to the gradational pattern, which is characteristic of Estonian common nouns, or to retain the original shape of the proper noun. For example, it has happened that within a single newspaper article the genitive case of the family name *Fink* appeared as the gradational form *Fingi* side by side with the non-gradational form *Finki*. It testifies that the author did not know which inflectional class to choose and, what may be even more important, the author was apparently unaware of the dilemma, as testified by the fact the article was published without harmonizing the inflectional confusion. It is only natural that in similar cases, difficulties would arise in automatic analysis, too.)

Suggestion: a morphology-conscious user dictionary

So a text normally contains words that have not been met in texts before, and are thus missing from the speller lexicon also, no matter how hard we have tried to improve it. How can we keep the speller from flagging the correct, although missing from the lexicon, words? An obvious way is to find some formal property that characterises unseen correct words: words with an uppercase letter, or containing a number, or looking like an internet or e-mail address. This is what MS speller option lets you define. According to our experiment with "Eesti Ekspress" (see above), this way one can diminish the number of false alarms by about 85%.

From what is left, we can treat some as productive newly-coined word forms, and thus pass them as correct (e.g. an English word ending with *–ly* may accept a suffix *–ness* as in *loneliness*). In Estonian, where it is common to create new words by gluing together previously known ones, the speller may pass as correct such compounds even if they are missing from the lexicon. (In the experiment with "Eesti Ekspress" such words were analyzed algorithmically, and thus were not counted as unrecognized).

This leaves us with a number of word forms (15% of all the un-recognized tokens, or 0.7% of all the word form tokens of "Eesti Ekspress"), some of which should be added to the user dictionary. This number may look small as an average, but it would be larger in case of specialised texts, and it is irritating for the user in any case. Here is where the user dictionary comes in handy… in principle.

We have received complaints about the user dictionary of the speller: why can the user add all the inflectional forms of a word to the dictionary only one by one? This is completely counter-intuitive for a native speaker of an inflectional language. It would be much more convenient if the speller guessed the inflectional class of a word, or, at least, allowed the user to choose from a predefined list an example word that the newly added word is similar to. From what we know about the nature of inflectional morphology, it would be possible to generate all the inflectional forms of the added word automatically (or at least to a great extent), and add them to the user dictionary.

It would truly diminish the number of false alarms by the speller: if a text introduces a new word, then in case of an inflective language, it is used in different word forms in the same text. At the moment, the speller API does not provide the possibility for adding this feature.

Conclusion

When maintaining the speller lexicon of an inflective language, it is possible to automate the process of defining the inflectional class of a previously unseen word.

The method is based on the notion of stable and unstable inflectional classes, which correspond to productive and un-productive classes, and to active and passive morphological rules.

A complementary feature of the predictability of the inflectional class of a new word is the observation that the size of the vocabulary of a language is infinite, and consequently, texts are bound to contain words that are missing from the speller's lexicon, no matter how big it is.

A practical solution would be to add a feature to the user dictionary that would make it possible to guess the inflectional class of the added word, generate all the inflectional forms of the word, and add them to the user dictionary. This would be clearly preferable to the current limitation that the inflected forms can be entered by the user only one by one.

References

1. Baayen, R. H. *Word Frequency Distributions*. Text, Speech and Language Technology, Vol 18, Kluwer Academic Publishers, Dordrecht/Boston/London 2001
2. Erelt, M., Kasik, R., Metslang, H., Rajandi, H., Ross, K., Saari, H., Tael, K., Vare, S. *Eesti keele grammatika. I Morfoloogia. Sõnamoodustus*. (*Grammar of the Estonian Language. I. Morphology. Word formation*) TA EKI, Tallinn 1995.
3. Viks, Ü. *Väike vormisõnastik* (*A Concise Morphological Dictionary of Estonian*), ETA KKI, Tallinn, 1992
4. Wurzel, W. U. *System-dependent morphological naturalness in inflection*. In: *Leitmotifs in Natural Morphology*, ed. W. U. Dressler. Studies in Language Companion Series, Vol. 10, John Benjamins, Amsterdam/Philadelphia, 1987, pp.59-95