Morphology-Aware Spell-Checking Dictionary for Esperanto

Marek Blahuš

Faculty of Informatics, Masaryk University
xblah@fi.muni.cz

Abstract. The article describes the process of constructing a spell checker for the Esperanto language and its implementation as a dictionary (i.e. an affix file and a word list) for the Hunspell spell-checking engine. In comparison to existing solutions, the chosen approach takes note of morphologically complex words, which are common in Esperanto due to its agglutinative nature, and applies a set of rules describing allowed morpheme compounds, along with semantic classification of all involved word roots. The result has been tested with a user community and is presently being incorporated into the *OpenOffice.org* office suite.

Key words: morphology; spelling; spell-checker; Esperanto

1 Introduction

The ease of electronic publishing is having a negative influence on the overall quality of texts, and the lack of accurate proofreading has had en especially grave impact on minority languages such as Esperanto [1]. Automated spell checking plays an essential role in helping the user produce quality texts.

There are several spell checking dictionaries for Esperanto [2], the most universal of which is that by Pokrovskij [3]. His solution, however, takes little note of Esperanto's rich morphology and is thus unable to recognize valid compounds such as "kaf·o·muel·il·o" ("coffee grinding machine") or "mal·sam·ras·an·oj" ("members of a different race").

In this paper, we describe a new spell checking dictionary for Esperanto, originally developed as a Bachelor thesis in the Natural Language Processing Centre at the Masaryk University [2]. Unlike the existing solution, it assigns each morpheme in the word list a set of semantic attributes and uses those in rules describing even complex Esperanto compounds. This has been made possible by the use of Hunspell [4], a modern spell-checking framework. Integration of the new dictionary in *OpenOffice.org* is also briefly discussed.

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2009, pp. 3–8, 2009.

2 Esperanto Morphology

Esperanto has an agglutinative morphology[1] based on *roots* and *lexical and grammatical affixes*.[2] The order of affixes around a root is important, since affixes modify the entire stem they are attached to. By means of compounding, stems may be joined together, either directly or with an *epenthetic* vowel.[3] Most words require at least one grammatical suffix on their end, by means of which part of speech and grammatical categories are expressed, but there are some roots that may lack it, such as the numeral "kvar" ("four"). Thus, structure of an Esperanto word may be described by the following regular expression:

$$(LexAfx^* \cdot Root \cdot LexAfx^* \cdot Epent?)^* \cdot LexAfx^* \cdot Root \cdot LexAfx^* \cdot GramAfx?$$

Evidently, not all strings matched by this expression are existing Esperanto words. The two following sections describe an attempt at eliminating the unexisting words. This is particularly important, as failing to do so would cause a significant drop in the spell checker's recall, since they often coincide with misspelings of existing words.

3 Word List

It has been suggested by prominent Esperanto linguists that every root has an inherent meaning, and that roots maybe grouped in classes according to their semantic characteristics. Wennergren [5, chapter 37.1] listed several such classes, such as *people, tools*, or *activities*, along with a couple of sample roots for each of them. He also pointed out that class membership of a root may have an influence on the set of possible word forming processes it can enter. This directly affects the productivity of affixes, as he indicates for instance in his description of the prefix "bo-" (parallel to the English suffix "-in-law") by stating that it may be used only with roots expressing family relationships.

Inspired by the classification sketched by Wennergren, we analyzed his descriptions of the behavior of all the 10 prefixes and 31 suffixes, and as to be able to fulfil the root class conditions imposed by each of them, we inferred a system which encompasses a total of 15 classes. They are shown in Table 1. Each root may member in a number of classes, but some classes are mutually exclusive (such as A, I and O) and some classes are actually subclasses of others (e.g. $F \subset P \subset O$). Altogether, there are 85 possible membership combinations.

Later on, we extracted 16,780 Esperanto roots from the electronic version of the PIV dictionary [6] and designed a system for their automatic semantic classification which determines the membership of each root in each class.

[1] With the exception of suffixes *-ĉj-* and *-nj-* used in affectionate forms of proper names and family relationships, whose presence has a truncating effect on the root, e.g. "patro" ("father") → "paĉjo" ("daddy").

[2] A constructed language, Esperanto has been designed so to decrease its user's memory load – by featuring affixes such as the prefix "mal-" for antonyms, e.g. "pez·a" ("heavy") → "mal·pez·a" ("light").

[3] This is being done due to euphony or if the inherent part of speech of the preceding root needs to be changed. Grammatical affixes *-o-*, *-a-*, *-i-* or *-e-* are used as such a link. See the compounds in Introduction for example.

Table 1. Semantic classification of roots

class	description
A	attribute roots, having the **a**-ending in their base word form
B	animals ("**b**estoj" in Esperanto)
C	**c**ommon gender in animals and persons
F	**f**emale gender in persons
I	action roots, having the **i**-ending in their base word form
J	place roots, producing adverbs of spatial meaning ("**ej**oj" means "places")
K	plants ("**k**reskaĵoj" in Esperanto)
L	antonym-producing roots, which accept the prefix "ma**l**-"
M	**m**ale gender in animals and persons
N	**n**umbers (numerals and several other roots expressing amount)
O	object roots, having the **o**-ending in their base word form
P	**p**ersons
T	**t**ransitive roots, producing transitive verbs
V	words which may appear without a grammatical suffix ("**v**ortetoj" means "little words")
Y	famil**y** relationships

Various linguistic resources such as corpora, specialized vocabularies and closed categories word lists are used in this step, some of which are listed in [2]. To search them and combine the results, a *Bash* script employing utilities from the *textutils* package has been used.

Sometimes, enumerating the roots that member in a class is not straightforward, as in the case of the L class. In such cases, corpus search for the prospective prefix-root combination has been conducted to prove or disprove its actual use. But as many valid words do not appear in corpora, probably not all members of the L class can be identified in this way. Other limitations, such as the insufficient size of some specialized vocabularies used, may also negatively influence the result.

4 Affix Rules

Within his ESPSOF project, Witkam [7] has produced a list of approximately 33,000 morpheme-segmentated Esperanto words, based on words appearing in the PIV and manually adjusted by him. We used this list to represent actual language use and inferred from it rules concerning allowed morpheme combinations within Esperanto words.

In the beginning, grammatical affixes were stripped and all roots within the words were identified, of which there may be several in a word, since Witkam's list includes also compounds. This has revealed that 39 % (12,970) of the words consist of a single root and no lexical affixes, 18 % (6,005) of them are a compound

of two roots and 4 % (1,386) are a compound of two roots linked by an epenthetic vowel "o". The remaining 38 % (12,639) are words containing lexical affixes, the most frequent *pattern* being a root followed by the "aĵ" suffix.[4] In total, 632 various word patterns have been discovered.

Later, all the roots were classified using the system described in the previous section, and every time all the words matching each of the discovered *patterns* (i.e. combinations of root placeholders and concrete lexical affixes) were examined at once. This examination has been done manually, using just common sense of a fluent Esperanto speaker, and its goal was to discover similarities among the roots that fell into the same place of the pattern, with the ultimate goal to replace this set of roots by a much smaller set of root classes. Based on the morphological assumptions made above, it should be possible to perform such an abstraction without excluding any existing words neither introducing words that do not exist (but it's probable that many existing words not present in Witkam's dictionary were included in this step).

Result of the automatic root classification and manually conducted examination and abstraction of the morphological patterns that emerged was a set of rules such as

[BKP]·"id"

meaning that the suffix "id"[5] may be attached after a root from either the B, K or P class, producing words like

"kat·id·o" ("kitten") from "kat·o" ("cat")

"kverk·id·o" ("oak offspring") from "kverk·o" ("oak")

"reĝ·id·o" ("prince") from "reĝ·o" ("king").

5 Implementation

In order to implement the designed spell checker as a dictionary (i.e. a word list and an affix file) for Hunspell, we had to find a workaround for Hunspell's very limited capabilities of working with regular expresssions. Currently, only the asterisk and the question mark operators are supported, of which only the question mark is of direct use for us – the optionary epenthetic vowel may be expressed by it. In the very frequent case when roots from several possible classes may occupy certain positions, the regular expression had to be split and separate expression had to be created, explicitly stating each of the possibilities.

This, along with word compounding, has led to dramatic increase of the number of regular expressions that form the affix file. Although some partial remedies have been found, such as grouping common sequences of classes into one virtual class of morpheme compounds, the resulting affix file has a size of

[4] This suffix is used to denote a concrete manifestation of the root, such as "manĝ·aĵ·o" ("meal") from "manĝ·i" ("to eat"). [5] This suffix is used to denote offspring or descendant of the root object, such as "hund·id·o" ("puppy") from "hund·o" ("dog").

37,155 rules, which slows down the spell-checking process, but fortunately not to any really remarkable extent yet. This could be avoided once a newer version of Hunspell would support the plus operator in its regular expressions.

The created Hunspell dictionary may be used for spell checking in software packages such as *Mozilla Firefox* or *OpenOffice.org*. For this, it needs to be provided with some additional information (such as meta information and license agreement) and packed up in form of an extension file for the particular application. Particular emphasis has been put on integrating the spell checker in *OpenOffice.org*, since a new Esperanto localization of this office suite is being prepared and the developed dictionary could become its official spell checker. For this, a dedicated subcomponent called "spellcheck" has been recently set up in *OpenOffice.org*'s bugtracking system *Issuetracker*, where users may submit their comments on the functionality of the dictionary.

6 Conclusion

In the article, we have described the process of developing a new spell-checking dictionary for Esperanto, with consideration of the language's word building system. We have developed a system of classes that reflect important semantic properties of word roots, as well as an automatic classification mechanism. We have inferred rules that make use of this class system to describe morphological structures of existing Esperanto words, and we have implemented these rules in form of a dictionary for the Hunspell framework.

Tests performed on computer transcription of an Esperanto-language literature manuscript[6] have shown a 19 % decrease (from 206 to 167) of misspelling recall in comparison with Pokrovskij's dictionary, which we consider an unpleasant side effect of extending the dictionaries morphological capabilities (what produces valid forms that coincide with misspellings) that, on the other hand, has caused a decrease in the amount of false positives (by 10 %, from 1565 to 546 unique words).

Future tasks include research on balancing the spell checker's precision and recall in order to achieve maximum user satisfaction, as well as further testing the developed solution with users and performing the necessary administrative steps to integrate the new dictionary as an automatically installed part of the Esperanto distribution of *OpenOffice.org*.

Acknowledgements. This study has been partially supported by the grants "Language helper for Esperanto" of the Esperantic Studies Foundation and MUNI33/212008 of the Faculty of Informatics of the Masaryk University and by the Ministry of Education of CR within the National Research Programme II project 2C06009.

[6] "Dívka na vdávání" by Miloslav Švandrlík, in Esperanto translation "Edzinigebla knabino" by Josef Vondroušek. 10,400 unique words. 52,700 words in total.

References

1. Petrović Lundberg, Sonja et al: Language helper for Esperanto – a project proposal. Esperantic Studies Foundation (2007).
2. Blahuš, Marek: A Spell Checker for Esperanto. Bachelor thesis, Faculty of Informatics, Masaryk University, Brno (2008).
3. Pokrovskij, Sergio: Vortaro por ISpell, 3.4 (2008) `http://www.esperanto.mv.ru/Download/ispell/ispelleo.tar.bz2`.
4. Németh, László: Hunspell : open source spell checking, stemming, morphological analysis and generation under GPL, LGPL or MPL licenses, 1.2.8 (2008) `http://hunspell.sourceforge.net/`.
5. Wennergren, Bertilo: Plena manlibro de Esperanta gramatiko. El Cerrito : ELNA, 2005. 696 pp. ISBN 0939785072. `http://bertilow.com/pmeg/`.
6. Plena ilustrita vortaro de Esperanto 2005. Editor Gaston Waringhien. Paris : SAT, 2005. 1265 pp. ISBN 2950243282.
7. Witkam, Toon: ESPSOF (Esperanto-Softvaro). Versio 0.8 (2008) `http://www.espsof.com/`.