Orwell's 1984 – Playing with Czech and Slovak Versions

Jaroslava Hlaváčová

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University, Prague
Malostranské nám. 25
hlava@ufal.mff.cuni.cz

Abstract. The contribution will describe an experiment with the automatic translational tool Česílko that was designed for translation between very close languages, namely Czech and Slovak. We had at our disposal the Czech version of the Orwell's novel 1984, morphologically annotated, and the Slovak version without the annotation. We automatically translated the Czech version into Slovak and compared the result with the automatic morphological annotation of the Slovak version. We evaluated the experiment using manually annotated part of the Slovak version.
During the experiment we had to deal with different inconsistent morphological tagsets used for the annotation of different input data. Conversions among them made the hardest problems of the whole work. The contribution will concentrate on overcoming inconsistent tagsets, as the problem is quite common.

1 Overview

The impulse for the experiment was a decision of our Slovak colleagues to use Orwell's novel *1984* for a manual morphological annotation. There exists the Czech counterpart — Czech translation of the same text morphologically annotated during the project MULTEXT-East that ran in 1995 – 1997 ([7]). It consisted in creating morphological tagsets for several languages of the Central and Eastern Europe and their use for annotation of the Orwell's novel.

The aim of the recent experiment was to preprocess the Slovak text by the automatic morphological annotation to speed the manual work of human annotators. Our idea was to try, if the automatic translator could help. We used our tools developed before, especially the automatic translator Česílko. With its help, we translated the Czech version automatically into Slovak. At the same time, we automatically morphologically annotated the Slovak version. Then we searched word forms from the original Slovak version in the translation, comparing their morphological tags. As the both Slovak texts (automatic and manual) differed, we had to develop a simple Slovak-Slovak aligner.

In the following sections, we will denote by ***C*** the Czech and by ***S*** the Slovak translation of the text *1984*.

2 Česílko – Translational Tool for Very Close Languages

Czech and Slovak are very close languages. So close, that an idea of a literal translation is not so crazy as it certainly is for the majority of other language pairs. The automatic translational tool Česílko takes advantage of the closeness. Its results were very promissing ([4]). There exists even its extended version for Lithuanian, with additional syntactic modules overcoming the greater distance between the languages ([3], [6]).

There are 2 possible ways, how to use the tool Česílko:

1. for translation between Czech and Slovak
2. for morphological annotation of Slovak texts

2.1 Translation from Czech to Slovak

The translation is based on morphological analysis of the intput text, in our case the Czech version. It assigns a set of pairs [lemma, morphological tag] to every word form. The set of pairs can contain tens of items for certain word paradigms, but the average is 3.6 pairs per word form. The basis for the morphological analysis is large monolingual morphological dictionary of Czech covering more than 800 000 lemmas (see [1]).

Next step is tagging, or disambiguation, and lemmatisation. It consists in choosing one pair [lemma, morphological tag] from the set of all possible ones. The tagging is based on statistics (see [2]) and achieves the accuracy between 92 and 93%.

For the translation itself, the bilingual Czech - Slovak dictionary is used, containing the data necessary for translation of lemmas and appropriate tags. It translates the Czech lemma assigned by the tagging and produces its concrete form in the second language (Slovak) according to the original (Czech) morphological tag selected by the tagger.

The translator is able to skip over the first two steps — morphological analysis and tagging — if the input text already contains a unique pair [lemma, morphological tag] for every word form. Thus, we could use our manually annotated data as the input.

2.2 Morphological Analysis and Tagging of Slovak

The morphological analysis of Slovak works identically as the morphological analysis for Czech described in the previous paragraph. It uses large monolingual Slovak morphological dictionary created by J. Hric covering more than 100 000 lemmas for the morphological analysis, and the same statistical methods for tagging.

3 The Data – Orwell's novel *1984*

For our experiments, we used 2 inputs:

1. the Czech version ***C***
2. the Slovak version ***S***

3.1 The Czech *1984* and its Pre-Processing

We had at our disposal the whole text of *1984* manually morphologically annotated, the result of the project MULTEXT-East ([8], [7]). As the morphological annotation of the Czech text was made manually, we can suppose that it was errorless. However, it was necessary to unify the tagsets. The original tagset (let us denote it ***TC1***) for the manual annotation was different from the tagset ***TC2*** (described in [1]) used by the translator and the both sets were not possible to transfer 1:1.

Namely, the tagset ***TC2*** uses compound values for several morphological categories, which is not the case ot the tagset ***TC1***. For instance masculine gender is not always further distinguished between animate and inanimate (especially for pronouns where there is often difficult or impossible to decide the right possibility) and has a compound tag for the both. Or, some morphological categories are possible to assign the value X, meaning "there can be any appropriate value". It is used for instance for the category of case with indeclinable nouns. The tagset ***TC1*** does not allow these possibilities. To make the both tagsets compatible, we had to exclude the detailed tags of ***TC1*** that did not have their counterpart in the tagset ***TC2***, and replace them with less detailed tags containing the compound values. Of course, we have lost some information.

The incompatibility between the tagsets ***TC1*** and ***TC2*** was the first main source of errors. The other was the "translation" of the Czech tagset ***TC2*** into the Slovak one ***TS1***, because of some differences, though tiny, between the both grammars. The tagset we used for automatic annotation of the Slovak version was created by J. Hric on the base of the Czech tagset of J. Hajič.

Later, we will mention the last source of errors, which is again connected with incompatible tagsets, this time the Slovak ones.

For the translation described in the previous section we used the annotated text ***C*** with converted tags, skipping the first two steps of the general procedure.

Let us denote ***CTS*** the Slovak translation of the Czech text (Czech Translated into Slovak).

3.2 The Slovak *1984*

The Slovak version of *1984* we used for our experiments was the "manual" translation of the novel, made by the human translator Juraj Vojtek [9].

The automatic morphological analysis of the Slovak text was processed, followed by the automatic statistical disambiguation (tagging). It consisted in choosing one pair [lemma, morphological tag] from all possible pairs proposed by the Slovak morphological analyzer. Let us denote ***SA*** the result of the automatic morphological analysis (Slovak Analyzed). Every word form now has assigned (possibly several) pairs [lemma, morphological tag], one of them automatically selected as the correct one.

There is an example of a single word form *piesku* (in English *sand* in genitive, dative or local) after this phase of procession:

<f>piesku<MDl>piesok<MDt>NNIS3-----A----<MMl>piesok<MMt>NNIS2-----A----<MMt>NNIS3-----A----<MMt>NNIS6-----A----

It has the following attributes assigned:

- <f> original word form, taken from the input text;
- <MMl> lemma assigned on the basis of the morphological dictionary; can be multiple;
- <MMt> tag assigned by the morphological analyzer; can be multiple;
- <MDl> lemma chosen by the tagger from the set created by the morphological analysis; always unique;
- <MDt> tag chosen by the tagger from the set created by the morphological analysis; always unique.

4 Slovak-Slovak Aligner

The first experiment consisted in comparing the Slovak texts ***CTS*** and ***SA***. As the Slovak translation ***SA*** from English was processed by an alive translator (a writer) and the translation ***CTS*** from Czech was automatic, there is no surprise that the both texts differ. We could neither align them on the basis of the same positions, nor it was possible to align sentences.

Table 1. shows the main numerical differences between the both texts:

	SA	*CTS*
# words	83 897	79 860
# delimiters	19 165	20 505
# sentences	6 974	6 714

However, as the both versions are in the same language, Slovak, there is no need to use complicated aligners designed for pairs of different languages. Our simple aligner consisted in trying to find the same word forms in the both texts. We did not search lemmas because in different translations they could appear in different grammatical forms which would not help us. If there appears the same word form in the both texts, there is bigger chance, that the both morphological tags could be the same, though there is quite great homonymy among the forms of one lemma. However, the homonymy usually exists only among one or two categories (see our example above, where the 3 MMt tags express the homonymy among genitive, dative and locative), the rest of the assigned tag could be right and could help a human annotator anyway.

The aligner worked in the direction ***SA*** $\rightarrow$ ***CTS***.

We looked within certain limits around the same position of the ***SA***. The limit (k) became the parameter of the aligner. As the texts have not the same number of positions, we had to assign a relative position to every word of the both texts. The relative position is a number from the interval $\langle 0, 1)$ expressing the relative location of the word within the whole text. It was calculated according to the following formula:

$$rel = \frac{1}{N} * abs$$

where N is number of words in the text, *abs* absolute position of the word expressing the order of the word — for the first word $abs = 0$, for the last one $abs = N - 1$. The relative position became the other attribute (rel) of our word forms:

<f>piesku<MDl>piesok<MDt>NNIS3-----A----<MMl>piesok<MMt>NNIS2-----A----<MMt>NNIS3-----A----<MMt>NNIS6-----A----<rel>0.00059597

To put it together, we sought in ***CTS*** every word form from ***SA*** within $\pm k$ words beginning on the nearest relative position in ***CTS***.

After preliminary experiments we found out that the most aligned word forms were prepositions and conjunctions. Because of their high frequency in any text, the alignment often made a pair of two items that did not belong to each other. That is why we excluded prepositions and conjunctions from our considerations.

If we found the same word form meeting the above conditions, we added a new attribute MPt to the word form in ***SA***:

<f>piesku<MDl>piesok<MDt>NNIS3-----A----<MMl>piesok<MMt>NNIS2-----A----<MMt>NNIS3-----A----<MMt>NNIS6-----A----<rel>0.00059597<MPt>NNIS3-----A----

In our example we see that the attributes MDt and MPt are equal. It means that the word form *piesku* from ***SA*** was found within our limits in ***CTS*** and the morphological tag chosen by the tagger is the same as that one assigned by the automatic translator.

The aligner sometimes found more than one identical word forms in ***CTS*** within the given span, especially for higher values of the parameter. These items had to be ignored because it is not possible to decide automatically, which is the right one. From the rest, more than 3/4 of the aligned word forms had the same morphological tags ($MDt = MPt$).

Table 2. shows results of this experiment for the parameters 10, 20, 30, 40, 50.

Parameter k	Identical words	More MPt's	$MDt = MPt$	% of unique alignment
10	3 829	681	2 514	79.86
20	6 268	1 827	3 496	78.72
30	7 973	2 912	3 877	76.61
40	9 297	3 894	4 125	76.35
50	10 942	4 714	4 765	76.51

It should be added, that there are 67 713 word forms that are neither prepositions, nor conjunctions. Even if the amount of equal tags is quite high, the automatic translation is not reliable enough to be used for assigning tags to a manual translation.

5 Evaluation

For the evaluation of the automatic morphological annotation we used the initial part of the novel *1984*, representing approximately one fifth of the whole text,

which had already been manually annotated with the new Slovak morphological tagset.

5.1 Differences Between Tagsets

Our Slovak colleagues decided to use again another tagset, different from the previous ones, and the incompatibility between the two systems represented a further source of errors. As the both systems of annotations are not possible to map 1:1, we had to adapt the conversion table in order that it could be used for comparing the manual and automatic results.

The biggest difference between the tagsets consists in annotating a special property of words — paradigm — by our Slovak colleagues. They distinguish up to 7 paradigms (nominal, adjectival, pronominal, numeral, adverbial, incomplete and mixed) for some parts of speech. This information is generally not possible to get from the Czech system of morphological tags. We had to ignore it and compare the results without the part of the Slovak morphological tag describing the paradigm.

Another big difference is Slovak distinguishing between adjectives and passive participia, even for already lexicalized items, which is not the case of the Czech system. In the evaluation we considered them equal.

Other incompatibilities were based again in possible compound values of the tagset ***TC2***, as has been described in the section 3.1. The Slovak system does not allow these possibilities.

We solved these problems by using simple regular expressions. Some morphological tags of the Czech tagset were translated into the Slovak format with some positions "dotted", so that they could be taken as regular expressions — one regular expression then could match with several Czech morphological tags. For instance the Czech tag NNFPX-----A---- for feminine (F) nouns (NN) in plural (P) with indeterminable case (X) was converted into SUfp. , where S means noun, U incomplete paradigm, f feminine, p plural and . (dot) stands at the case position. It is not part of the Slovak tagset, but the whole tag can be used as a regular expression with an arbitrary sign at the end.

Having settled up the problems with different tagsets, we converted Czech morphological tags MDt and MPt into Slovak SDt and SPt. We also added the tag MAN from the manual annotation, so that the final word forms looked like:

<f>piesku<MDl>piesok<SDt>SSis3<SPt>SSis3<MAN>SSis3

The following example has a tag in the form of regular expression. The dot stands at the position of the paradigm, that was not possible to get from the Czech system. The rest of the manual tag is the same; we could evaluate such cases as successful.

<f>jeden<MDl>jeden<SDt>N.is1<SPt>N.is1<MAN>NFis1

5.2 Results

We introduced the last attribute AGR (agreement). It can have the following values:

D , if $MAN = SDt$ (agreement between the manual annotation and Slovak annotation)
P , if $MAN \neq SDt$ and $MAN = SPt$ (agreement between the manual annotation and the translation)
0 , otherwise (no agreement)

The table 3 shows the final results for word forms without delimiters:

AGR	Count	%	Explanation
D	14 226	80.35	$SDt = MAN$
P	86	0.49	$SPt = MAN \& SDt \neq MAN$
0	3 392	19.16	no agreement

The reason of not very high agreement lies in the multiple conversions among incompatible tagsets.

Let us have a look at items of no agreement. Almost one quarter of them (22.8%) are unknown words (SDt = Q) — words that were not present in the morphological, nor in the translational dictionary. Half of them are proper names (for instance *Winston* has the frequency 544). There is always problem with proper names because there will never exist a dictionary that would contain them all. However, it is possible to recognize them with other methods, for instance guessers ([5]).

Many errors are caused by the incompleteness of the dictionaries. One of the useful results was the list of words that should be added. However, due to the subject matter of the novel, there is quite a lot of very unusual words that are not used in the current language — *ideozločinec* (in English *thought-criminal*), *podpododdelení* (in English *subsubdivision*), some of them were even not translated into Slovak — *newspeak*, *facecrime*, or have a special Slovak ending — *goldsteinizmus*. These types of words do not belong to general dictionaries, it is necessary to recognize and determine them by different means (an automatic recognition tool, a guesser).

5.3 Conclusions

Though the automatic translational tool Česílko itself is reported to be very good, it is not possible to use it directly for annotation of the original Slovak text. However, it is possible to align the manual and automatic texts very easily on the basis of individual word forms.

Different approaches to basic inputs bring a lot of hardly surpassable barriers that are necessary to overcome at the cost of losing accuracy. For better results, it would be necessary to "translate" the dictionaries into the final tagset. Unfortunately, it cannot be done entirely automatically — it would demand a lot of manual work.

Acknowledgements

The work reported in this paper arose from the Project of Scientific and Technical Collaboration of the Czech Ministry of Education with Slovakia, number 150. It has also been supported by the grants of the Czech Ministry of Education 1ET101120503 and 1ET101120413.

References

1. Hajič, J.: Disambiguation of Rich Inflection. (Computational Morphology of Czech) Praha, Karolinum 2004
2. Hajič, J.; Hladká, B.: Tagging Inflective Languages. Prediction of Morphological Categories for a Rich, Structured Tagset. ACL-Colint'98. Montreal, Canada, August 1998.pp.483–490
3. Hajič, J.; Homola, P.; Kuboň, V.: A Simple Multilingual Machine Translation System. In Proceedings of Machine Translation Summit 2003 IX, pp. 157–164.
4. Hajič, J.; Hric, J.; Kuboň, V.: Machine Translation of Very Close Languages. Proceedings of the ANLP 2000, Seattle, USA, April 2000, pp. 7–12.
5. Hlaváčová, J.: Morphological Guesser of Czech Words. Proc. TSD 2001. Springer-Verlag Berlin Heidelberg 2001, pp. 70-75.
6. Homola, P.; Kuboň, V.: A translation model for languages of acceding countries. In Proceedings of the EAMT Workshop 2004
7. MULTEXT-East project: `http://nl.ijs.si/ME/`
8. Petkevič, V.: Czech translation of G. Orwell's '1984': Morphology and syntactic patterns in the corpus. Number 1692 in Lecture Notes in Artificial Intelligence, pages 77-82. Springer-Verlag, 1999.
9. Orwell George: 1984. Translated by Vojtek J. Bratislava, Slovart 1998. ISBN 80-7145-334-X