CzAccent – Simple Tool for Restoring Accents in Czech Texts

Pavel Rychlý

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
pary@fi.muni.cz

Abstract. There are many Czech text written without any accents. The paper describes a tool for fully automatic restoration of Czech accents. The system is based on a simple approach of big lexicon. The resulting accuracy of the system evaluated on large Czech corpora is quite high. The system is in regular use by hundreds of users from around the whole world.

Key words: diacritic restoration, Czech, CzAccent

1 Introduction

The written form of the Czech language uses the same 26 character as English many of them with several different accents. The list of all accented characters of Czech is in Table 1. In the early days of personal computers in 80s of the last century and in the early days of mobile phones in the beginning of this century, the devices was to prepared for easily writing of accented characters and users wrote Czech texts without accents. Even today, there are people who write some texts or all texts without accents.

For most people, reading Czech text without accents is harder than reading correct texts. Many non-accented words are ambiguous, there are more than one possible way how to add one or more accents to create different words. On the other hand, all native speakers do not have problems with understanding the non-accented text, they are able to add correct accents to words in a context.

Table 1. All accented characters in Czech.

á	í	ť	Á	Í	Ť
č	ň	ú	Č	Ň	Ú
ď	ó	ů	Ď	Ó	Ů
é	ř	ý	É	Ř	Ý
ě	š	ž	Ě	Š	Ž

Aleš Horák, Pavel Rychlý (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2012, pp. 85–89, 2012.

Table 2. Relative frequency of accented characters in Czech texts compared to respective non-accented ones.

character	% in text	character	% in text
a	6.861	r	4.002
á	2.073	ř	1.129
c	2.543	s	4.617
č	0.962	š	0.779
d	3.659	t	5.583
ď	0.024	ť	0.043
e	7.996	u	3.165
é	1.077	ú	0.142
ě	1.424	ů	0.496
i	4.617	y	1.732
í	2.896	ý	0.855
n	6.517	z	2.020
ň	0.066	ž	0.972
o	8.146		
ó	0.030		

There was several attempts to build an automated tool for adding accents, usually based on learning n-grams of characters. The presented system outperform all character based systems.

The structure of the pager is following: next section states the complexity of the problem, then the CzAccent system is described in details. Next two sections provide results of evaluation and usage options of the system.

2 Complexity of Restoring Czech Accents

Accented vowels are very common in Czech texts, many other accented characters are very rare. The relative frequency of accented characters together with respective non-accented variant are listed in Table 2.

We can see that most frequent accented characters *á* and *í* are also most frequent accents compared to respective non-accented characters. The *á* character occurs in 23 % of all *a* occurrences (accented or non-accented). The *í* character occurs in almost 40 % of all *i* occurrences.

3 CzAccent method

The CzAccent system is based on a big lexicon. The the all words known to the Czech morphology analyser Majka was looked in a big corpus. The most frequent accented word from all possible accented words and also the original non-accented word was selected and added to the CzAccent lexicon. In the

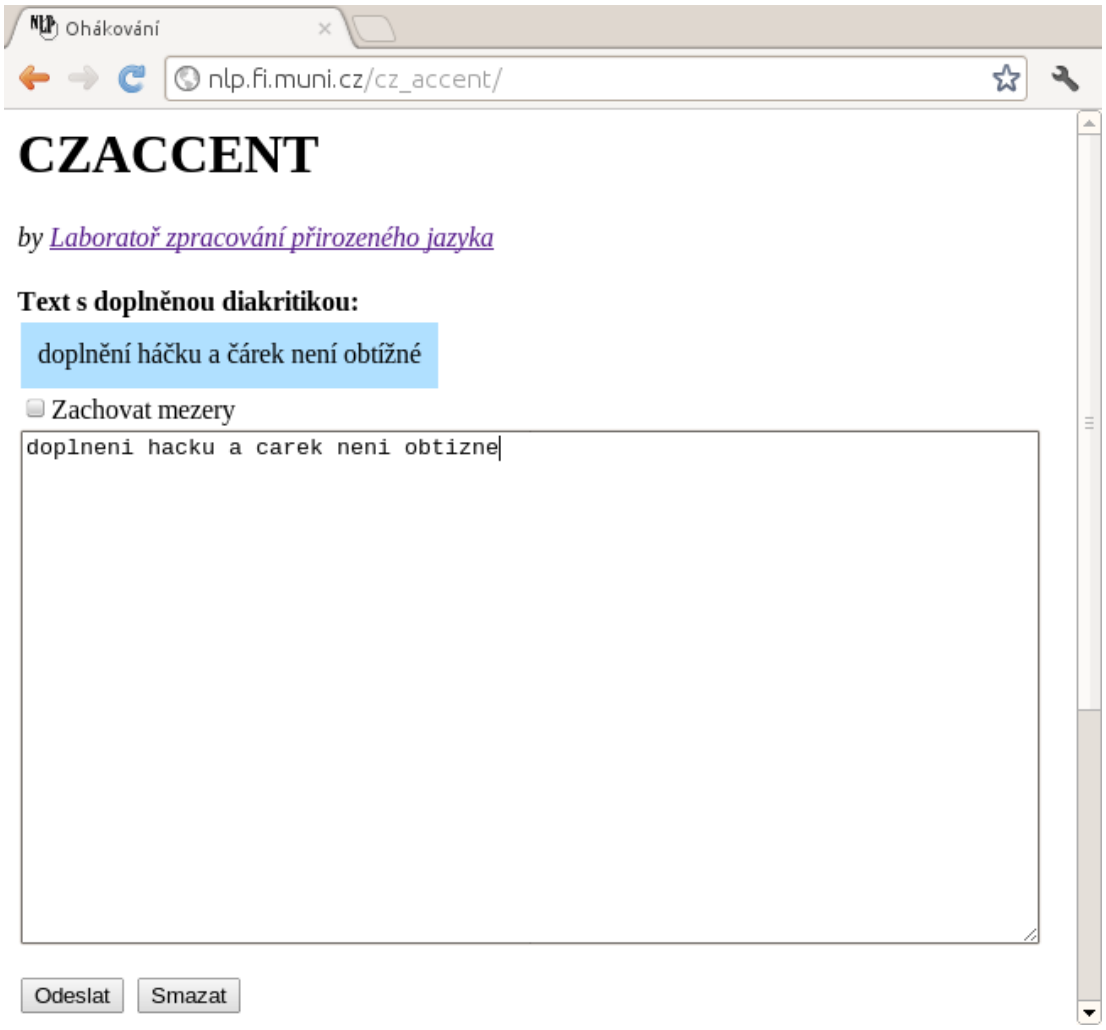

Fig. 1. CzAccent web interface.

result, there are millions of words which are stored in a very compact data structure using a finite state automaton [1]. The final data file containing the whole lexicon is very small, it has only 1.86 kB.

The CzAccent system processes any text in a straightforward way, it repeats the following steps from the beginning to the end of an input.

1. read a word (sequence of alphabetical characters),
2. try to find the word in the lexicon,
3. if found print the accented variant,
4. else print original word,
5. copy any non-alphabetical characters from input to output.

Due to its simplicity, the system is very fast. It can process more than 4.7 MB per second on moderate machine.

4 Evaluation

The system was evaluated on big Czech corpus CZES. CZES was built purely from electronic sources by mostly automated scripts and systems. [2]
Texts in the CZES corpus come from three different sources:

1. automated harvesting of newspapers (either electronic version of paper ones or electronic only), with annotation of publishing dates, authors and domain; these information is usually hard to find automatically from other sources;
2. customised processing of electronic versions of Czech books available online; and
3. general crawling of the Web.

The whole corpus should contain Czech texts only. There are small parts (paragraphs) in Slovak or English because they are parts of the Czech texts. Some Czech newspapers regularly publish Slovak articles, but we have used an automatic method to identify such articles and remove them from the corpus.

There was no restriction on the publication date of texts. There are both latest articles from current newspapers and 80 year old books present in the corpus.

The second corpus for evaluation was 1 million word corpus DESAM [3]. It is manually annotated and that is the reason that it is also very clean.

The accuracy of the system on the CZES corpus is 92.9, the accuracy on DESAM is 97.3. We can see that on cleaned texts the accuracy is very high.

5 Interface

The system is stable, it can be run in the form of a command line tool. An example of usage is at Figure 2.

```
$ echo realny problem | czaccent
reálný problém
```

Fig. 2. An example of CzAccent command line tool usage.

There is also a public web interface at the following address: `http://nlp.fi.muni.cz/cz_accent/`. It is in the form a simple page with one entry field. A user can enter a Czech text without accents and the system provides accented text. A screen-shot of the web interface is at Figure 1.

6 Conclusions

The system uses very simple method, but the resulting accuracy of the system evaluated on very large Czech corpus is quite high. The system is in regular use by hundreds of users from around the whole world.

Acknowledgements

This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013 and by EC FP7 project ICT-248307.

References

1. Daciuk, J.: Finite state tools for natural language processing. In: Proceedings of the COLING 2000 workshop Using Toolsets and Architectures to Build NLP Systems, Luxembourg. (2000)
2. Horák, A., Rychlý, P.: Discovering grammatical relations in czech sentences. (2009)
3. Pala, K., Rychlý, P., Smrž, P.: DESAM – Annotated Corpus for Czech. In: Proceedings of SOFSEM '97, Springer-Verlag (1997) 523–530