MorphCon – A Software for Conversion of Czech Morphological Tagsets

Petr Pořízka[1] and Markus Schäfer[2]

[1]Department of Czech Studies, Faculty of Arts,
Palacký University in Olomouc, Czech Republic
petr.porizka@upol.cz
[2]Institute of Computer Science
The University of Bonn, Germany
schaefe1@informatik.uni-bonn.de

Abstract. This study reflects current situation in Czech corpus linguistics with a special view to morphological annotation of language corpora. Several morphological tagsets of Czech exist nowadays. These tagsets differ by the conception reflecting morphological categories in different extent of complexity. There has also been no possibility of conversion among tagsets. New tool called *MorphCon* (Morphological Convertor) is now being developed for these purposes. This first version (0.1alpha) enables converting of two basic morphological tagsets of Czech: Prague positional system and Brno's attributive system. There are three basic Input/Output (I/O) formats of data (SimpleTag-Conversion, KWIC/Tag-Format, WPL-Format) within version 0.1alpha. Tagsets are implemented into the *MorphCon* as "drivers" with "encode" and "decode" function as well as an "universal library" called *DZ-Interset* (© Daniel Zeman) – modified in our tool – plays key role for the process of conversion as a transcoder. The *MorphCon* software is thus built as an universal converter: modularity, the *Interset* as a transcoder, possibility of adding of another tagsets (not only Czech ones) and I/O formats.

1 Introduction

The development of the *MorphCon* software (shortened from ***Morph****ological* ***Con****vertor*), an application based on formal rules and algorithms, was motivated by the current situation of Czech corpus linguistics with respect to morphological annotation of linguistic corpora. At present several morphological tagsets exist for Czech, which reflect morphological categories to various degrees of complexity and which differ one from another in their conceptions. The predominant and most used system is the one designed by J. Hajič (a positional tagset, hereafter, PT) [6][7], as seen in the written part of the *Czech National Corpus* [18]. Not less important is the Brno system of morphological tags (an attributive tagset, hereafter, AT), the conception of which is authored by K. Osolsobě [11] and R. Sedláček [17]. This system is made use of by the morphological analyzer (tagger) called AJKA [15][16] in the corpora of the *Natural Language Processing Centre* at Masaryk University, Brno (hereafter, NLP FI MU).

Among others belongs Petkevič´s morphological tagset used in the international project MULTEXT-EAST (the Orwell 1984 Corpus) [13] or, most recently, the so-called *kódovník* (coder), used lately for tagging of the *Prague Spoken Corpus* [4].

The tagsets mutually differ with respect to their conception, the degree of complexity and system as far as the coded grammatical categories are concerned. Morphological tagsets also depend on particular software applications, which are not composed to annotate linguistic data by means of another tagset, as they operate with one particular tagset only. However, if necessary, it is possible to annotate the same text by various tagsets by means of the respective applications, but the crucial role is played here by the efficacy, i.e. the degree of success and failure at (semi)automatic processing of the texts by the given taggers, the scope and structuring of their dictionary for lemmatization and tagging etc.

One of the possible solutions to this situation is a software allowing automatic conversion of (Czech) morphological tagging systems. For these reasons, the *MorphCon* converter has been developed, which allows the conversion of corpus data already tagged by one tagset into a different tagset.

The conversion between the Prague and Brno tagsets has been chosen for the first phase of this software. The Prague system is a positional one (15 positions are given, to each of which a particular linguistic category is assigned, represented by a defined subset of symbols), the Brno system is an attributive one (there is a combination of diagrams, in which the first symbol represents a grammatical category, the second one its concrete value for the word given). A detailed account and an overview of the symbols of both systems are presented in Hajič [6][7], Osolsobě [11], and Sedláček [15][16][17]. For us to find out the possibilities of the mutual automatic conversion, it was necessary to carry out an analysis and comparison of both tagsets, the key factors being mutual compatibility, loss of information and other aspects (see below).

2 The *MorphCon* converter

The software *MorphCon* (v0.1alpha) is in development since 2008 by a team of authors consisting of members of three universities: Petr Pořízka (Faculty of Arts, Palacký University, Olomouc), Marek Schäfer (Faculty of Informatics, Bonn University, Germany) and Daniel Zeman (Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague). It is developed in the *Perl* scripting language (v5.10.0) and designed as a universal converter, the fundamental elements of which are modularity and existence of a conversional tagset (which excludes the direct conversion). This allows implementation of further tagsets and Input/Output formats.

With respect to possible future users-linguists, the so-called Graphic User Interface (GUI) has been kept in mind.

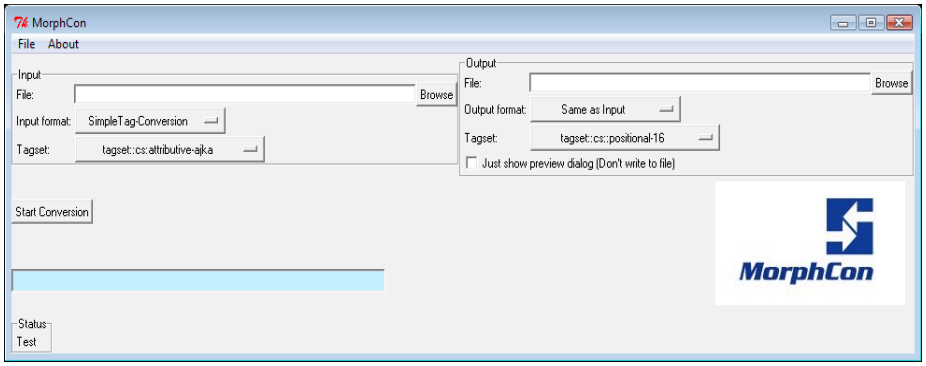

Fig.1. The Graphic User Interface of the MorphCon converter

2.1 The structure of the software

The conception, structure and the principles of the *MorphCon* converter are presented in the *Scheme* in the Supplement, to which the following text refers to. The *MorphCon* is composed of several components/modules (all written in the Perl scripting language):

- GUI: the Graphic User Interface – MorphCon.pl
- Input/Output modules: MorphCon::{simple,kwic,wpl}
- Drivers: implemented morphological tagsets – tagset::cs::{attributive-ajka,pdt,positional-16}
- Universal library: a modified *DZ Interset* library

The key quality is modularity of the software, apart from offering universality (including the possibility to enrich the *MorphCon* with other tagsets) also variability in the process of conversion, i.e. various/different settings of the input and output data. The *MorphCon* is based on the universal tagset *DZ Interset* (see [21] [22]) which works as a transducer when converting one tagset into another. Each tagset is implemented into the software as a „driver“ with double function, either as a source, or target tagset:

- encode-function: source-tagset → Interset
- decode-function: Interset → target-tagset

2.1.1 The Interset

The *Interset* functions as a „feature-projection“, i.e. it is structured as a „feature::value“ system, the grammatical, morphosyntactic category being the feature. It must also contain all the features (grammatical categories) with their values from all implemented tagsets. An overview of all features and drivers may be found on the websites of the project [24]. During the conversion, the tags from the tagset A

are converted into the *Interset*, from which they are subsequently converted into the target tags within the tagset B. The quality of conversion thus depends on the quality of algorithms converting the particular categories with values of the tagsets given into the "feature::value" system of the *DZ Interset* [23].

Example: The *Interset* and the "feature::value" structure for the attributive tag 'k2eAgInPc6d1' from the ajka-tagset in the Perl script:

```
{
 'animateness' => 'inan',
 'case' => 'loc',
 'degree' => 'pos',
 'gender' => 'masc',
 'negativeness' => 'pos',
 'number' => 'plu',
 'pos' => 'adj',
 'possgender' => 'masc',
 'possnumber' => 'plu',
 'tagset' => 'cs::ajka'
}
```

	Annotation of the tag '**k2eAgInPc6d1**'	
k2	k = part of speech	2 = adjective
eA	e = negation	A = affirmation
gI	g = gender	I = masculine inanimate
nP	n = number	P = plural
c6	c = case	6 = sixth
d1	d = degree	1 = positive

The principle of conversion of this tag from an attributive one into a positional tagset with representation in the Perl script in the "AT,PT {Interset} format" looks as follows:

Example:

```
k2eAgInPc6d1,AAIP6MP--1A-----{'animateness' => 'inan','case'
=> 'loc','degree' => 'pos','gender' => 'masc','negativeness'
=> 'pos','number' => 'plu','pos' => 'adj','possgender' =>
'masc','possnumber' => 'plu','tagset' => 'cs::attributive-
ajka'}
```

2.1.2 The input/output modules

The Input/Output modules of the *MorphCon* noticeably extend the options of the *DZ Interset*, since – apart from the process of conversion itself – they allow to set the input and output data variably. Thus, the input and output formats do not have to be identical.

The options for the Input/Output modules

- **the data file**
  - the *.txt* format, "plain text"
- **the tag format**
  - the Brno tagset – attributive: tagset::cs::attributive-ajka
  - the Prague tagset – positional: tagset::cs::pdt
    tagset::cs::positional-16
- **the file format**
  - simple: SimpleTag-Conversion
  - KWIC: KWIC/Tag-Format
  - WPL: WPL-Format

Note on the data file

So far, the *MorphCon* has only been working with the plain text (*.txt* files), but in the future it should also process texts with structured data format (SGML and especially XML data).

Note on the tag format

There are two variants for the positional tagset: (a) "tagset::cs::pdt" is the original driver of the *Interset*, in which the category of aspect is not included; (b) "tagset::cs::positional-16" comes from the original driver, but the important category of aspect has been newly added to it in the *MorphCon*. (This means an extension from 15 to 16 tagset positions.) Therefore the positional system of tags may be enriched by means of conversion in the direction AT → PT, as the attributive system counts with the aspect category, while the positional system does not (apart from some exceptions). It is a way to prevent a loss of information during conversion in such a case when the source tagset contains a grammatical category not found in the target tagset.

Note on the file format

As mentioned above, the *MorphCon* allows the change of input and output file formats. When converting tagset A into tagset B, it is thus possible to determine the resulting data format, differing from the source data format A. So far the *MorphCon* has been working with three formats: (1) The SimpleTag-Conversion (shortened: simple), (2) The KWIC/Tag-Format (shortened: KWIC), and (3) The WPL-Format (shortened: WPL). See the following examples of these formats:

2.1.2.1 The SimpleTag-Conversion

This format is a simple conversion from tagset A into tagset B. One separate line corresponds to one tag of the respective tagset:

AT	PT
k3gInPc6	PDXP6----------
k2eAgInPc6d1	AAFP6----1A-----
k1gInPc6	NNIP6-----A-----

Table 1. The simple "tag-to-tag" conversion

2.1.2.2 The KWIC/Tag-Format

The KWIC/Tag-Format allows the conversion of tags from already annotated and completed corpora. In practice, the so-called concordances are most frequently searched for, i.e. search results where the key word (KWIC = Key Word In Context) is surrounded by a context (see the example taken from the *Bonito* corpus manager – for more detailed information about the *Bonito* concordancer see Rychlý [14]). Usually the tag is placed as a metatext following the word, this means exactly in the format KWIC/Tag (the slash is a division mark). In the converting process, only tags are converted, the remaining text is kept unchanged.

The source tagset (AT) – the pre-conversion phase

172 měl zajištěnou nominaci na Světový pohár v < Petrohradu /k1gInSc6wS > . " Má Švanda naději startovat

284 prezident Václav Havel za přítomnosti dalších < hostí /k1gMnPc2wK > . Za zvuku státní hymny vztyčili

↓

The target tagset (PT) – the post-conversion phase

172 měl zajištěnou nominaci na Světový pohár v < Petrohradu /NNIS2-----A----- > . " Má Švanda naději startovat

284 prezident Václav Havel za přítomnosti dalších < hostí /NNMP2-----A---2- > . Za zvuku státní hymny vztyčili

2.1.2.3 The WPL-Format

The corpus data are usually structured into the so-called vertical format, there is always only one word (WPL = Word Per Line) with its linguistic interpretation per line. For that reason, the sequence *word – lemma – tag* is most often separated by a tabulator (comma or other marks are acceptable, too). The WPL-Format format is therefore very important in case we build a corpus: we annotate the data with the help of tagset A, but we may subsequently need to convert the data into tagset B. The reason for this may be situations when a tagger with implemented tagset A has at its disposal a higher-quality and larger dictionary and gives higher-quality results (i.e. proportionally fewer errors during the process of automatic annotation) than a tagger with implemented tagset B. In comparison with tagset A, tagset B is more user-friendly, may be remembered more easily etc. There may be a whole variety of reasons for using this format.

AT: word	lemma	tag	PT: word	lemma	tag
V	v	k7	V	v	RR--6----------
těch	ten	k3gInPc6	těch	ten	PDXP6----------
dlouhých	dlouhý	k2eAgInPc6d1	dlouhých	dlouhý	AAFP6----1A-----
rozhovorech	rozhovor	k1gInPc6	rozhovorech	rozhovor	NNIP6-----A-----

Table 2. The sequence word – lemma – tag in the WPL-Format

3 Different conceptions of AT and PT systems

Important facts for conversion, i.e. conversion algorithms, are the question of different conceptions of tagsets and their mutual convertibility, the potential loss of information during conversion, and other aspects. There may be cases where modifications of tagsets are concerned, see the example of a positional tagset supplied with another position representing the category of aspect.

Problems when converting the attributive and positional tagsets of Czech appeared, in fact, we still find ourselves at the stage of testing and adjusting the conversion algorithms to optimize the *MorphCon* conversion process.

The positional tagset is characterized by the controversial second position (SUBPOS), which is a rather unsystematic and heterogeneous mixture of symbols from various grammatical categories of various parts of speech and with various degrees of complexity (cf. [7]). The second position contains 75 values altogether, which according to our opinion could be reduced, avoiding loss of relevant linguistic information. The „double-category" of the 3rd/6th and 4th/7th positions (gender and number), and existence of the 6th and 7th categories as separate positions expressing possessive gender and possessive number, respectively, are also debatable. On the other hand, the grammatically important category of aspect is missing. The category of aspect has been indeed added as the 16th position to the annotated corpora of Czech called *SYN2005* and *SYN2006PUB* (cf. [2][3]), but it has not been implemented into the positional tagset (and tagger). The question remains, what the still unexploited „reserved" positions 13 and 14 are intended for, and why the tagset still contains these two empty positions. The attributive tagset is more complex concerning the contained linguistic (sub)categories and could be characterized as the one that is more systematic regarding the conception, more user-friendly regarding acquisition and more easily remembered. The problem that appeared when working on the conversion algorithms concerned documentation and a complete set of tags. There exist several versions of the attributive tagset from various development stages, but there is no version history available, and therefore we had to take all of them into account. There exist four versions of the attributive tagset in total: (1) the official version by R. Sedláček (the author of *Ajka*, the tagger using the attributive tagset), available on the website of NLP FI MU [17]; (2) the tagset table by R. Sedláček from 2006; (3) the tagset version found in the disambiguation manual by Bartůšková et al. [1]; (4) the original tagset version from R. Sedláček´s diploma thesis [15].

A number of Perl scripts were created for testing of the *MorphCon*, which had partial functions during the conversion process, e.g. indicating tags of the same content and visualization of differences between "encode" and "decode" functions of "drivers", i.e. tagsets. With help of testing scripts, the problematic situations and errors originating before the conversion, namely already during the process of annotation of corpus data, may be detected. Mainly non-uniformed tag order, erroneous annotations of particular tokens, which remained in the text most probably due to insufficient disambiguation or were inserted into the tags by a disambiguator (the „human factor“), e.g. an attribute without a value or a value without an attribute, more values in the category of case and other errors.

4 Further advancement of the *MorphCon*

In the near future, we plan to expand functionality of the *MorphCon*. We intend to implement other tagsets of Czech (the above mentioned *mte-cz tagset* or *kódovník*), further Input/Output formats (both linear and vertical), e.g. the format used by the *Ajka* tagger [15] or in the *Prague Dependency Treebank* (the so-called csts /pml/ format – for more details, see *PDT 2.0 Guide*, Chapter 3. Data) [12].

Following the experience acquired during the development of the *MorphCon* and solving problems relating to the conversion, implementation of some scripts originally intended for testing purposes only were considered. The aim is to produce a new module „Tag-Checker“, which would serve as a helping instrument for disambiguation and tag checking (checking their correctness/error rate, anomalies etc.). Basic information about the *MorphCon* software may be found at http://www.morphcon.webnode.cz.

References

[1] Bartůšková, D., Hlaváčková, D., Ungermanová, M. (2004). *Manuál pro značkování a desambiguaci slovních tvarů v jazykových korpusech*. Brno. [online], [cit. 2009-05-29]. Available from: <http://nlp.fi.muni.cz/projekty/desman/desman1603.pdf>

[2] *Corpus SYN2005*. [online], [cit. 2009-05-29]. Available from: <http://www.korpus.cz/english/syn2005.php>

[3] *Corpus SYN2006PUB*. [online], [cit. 2009-05-29]. Available from: <http://www.korpus.cz/english/syn2006pub.php>

[4] Čermák, F. et al. (2007). *Frekvenční slovník mluvené češtiny*. Praha: Karolinum.

[5] Hajič, J., Hladká, B. (1998). Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structured Tagset, In: *Proceedings of the Conference COLING – ACL `98*. Montreal, Canada, pp. 483–490.

[6] Hajič, J. (2005a). Popis morfologických značek – poziční systém. In: *Manuál korpusového manažeru Bonito*. [online], [cit. 2009-05-29]. Available from: <http://www.korpus.cz/bonito/znacky.php>

[7] Hajič, J. et al. (2005b). *A Manual for Morphological Annotation – Positional Tags*. [online], [cit. 2009-05-29]. Available from: <http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/m-layer/html/ch02s02s01.html>

[8] Hladká, B. (2000). *Czech Language Tagging*. PhD Thesis. Charles University, Prague.

[9] *MorphCon – Convertor of Czech Morphological Tagsets*. [online], [cit. 2009-05-29]. Available from: >http://morphcon.webnode.cz>

[10] *Natural Language Processing Centre* (Faculty of Informatics, Masaryk University, Brno). [online], [cit. 2009-05-29]. Available from: <http://nlp.fi.muni.cz/en/nlplab>

[11] Osolsobě, K. (1996). *Algoritmický popis české formální morfologie a strojový slovník češtiny* (unpublished PhD Thesis). Brno.

[12] *PDT 2.0 Guide*. [online], [cit. 2009-05-29]. Available from: <http://ufal.mff.cuni.cz/pdt2.0/doc/pdt-guide/en/html/index.html>

[13] Petkevič, V. *Popis morfologických značek použitých v korpusu orw-mte*. [online], [cit. 2009-05-29]. Available from: <http://ucnk.ff.cuni.cz/orwell_znacky.php>

[14] Rychlý, P. (2000). *Korpusové manažery a jejich efektivní implementace*. (PhD Thesis) Brno. [online], [cit. 2009-05-29]. Available from: <http://www.fi.muni.cz/~pary/dis.pdf>

[15] Sedláček, R. (1999). *Morfologický analyzátor češtiny*. (MA Thesis) Brno. [online], [cit. 2009-05-29]. Available from: <http://nlp.fi.muni.cz/projekty/ajka/ajka.pdf>

[16] Sedláček, R. (2004). *Morfologický analyzátor češtiny Ajka.* [online], [cit. 2009-05-29]. Available from: <http://nlp.fi.muni.cz/projekty/ajka/ajkacz.htm>

[17] Sedláček, R. (2006). *AJKA tagset*. Brno. [online], [cit. 2009-05-29]. Available from: <http://nlp.fi.muni.cz/projekty/ajka/tags.pdf >

[18] *The Czech National Corpus*. [online], [cit. 2009-05-29]. Available from: <http://www.korpus.cz/english/index.php>

[19] *The Prague Dependency Treebank*. [online], [cit. 2009-05-29]. Available from: <http://ufal.mff.cuni.cz/pdt2.0/>

[20] *ÚFAL Tools* (Institute of Formal and Applied Linguistics at the Faculty of Mathematics and Physics, Charles University, Prague). [online], [cit. 2009-05-29]. Available from: <http://ufal.mff.cuni.cz/tools.html>

[21] Zeman, D. (2008). Reusable Tagset Conversion Using Tagset Drivers. In *Proceedings of the Language Resources and Evaluation Conference, LREC 2008*. CD full edition + printed Conference Abstracts. Marrakech, Morocco. [online], [cit. 2009-05-29]. Available from: <http://ufal.mff.cuni.cz/~zeman/publikace/2008-02/tagdrivers-marrakech-styl-lrec.pdf>

[22] Zeman, D. *DZ Interset*. [online], [cit. 2009-05-29]. Available from: <https://wiki.ufal.ms.mff.cuni.cz/user:zeman:interset>

[23] Zeman, D. *DZ Interset – Features*. [online], [cit. 2009-05-29]. Available from: <https://wiki.ufal.ms.mff.cuni.cz/user:zeman:interset:features>

[24] Zeman, D. *DZ Interset – Tag Set Drivers*. [online], [cit. 2009-05-29]. Available from: <https://wiki.ufal.ms.mff.cuni.cz/user:zeman:interset:drivers>