Conversion of *Morfologik* data into the *IKIPI* tagset

Adam Radziszewski
Institute of Informatics, Wrocław University of Technology
27.01.2011

This document describes some technical details of tagset conversion that was performed on the data from *Morfologik* dictionary (morfologik.blogspot.com). The conversion was performed semi-automatically, resulting in most of the source entries compliant with the IKIPI tagset.

Input data and output tagset

The input data is stored in a three-column tab-separated text file. Subsequent columns correspond to orthographic form, lemma and *tag representation*. Tag representation may stand for single tag or multiple tags. In the latter case, several tags are writtern as a notation shorthand using dot and/or underscore characters. We used the Maca system (and accompanying Python scripts) to read, process, compact and write this format [MACA].

Originally, the data were taken from Morfologik 1.7 RC2. Next (25.10.2010), the differences between 1.7 RC2 and 1.7 final were accounted for (there were not many of them), so the dataset labelled `1222` (as of 22.12.2010) reflects Morfologik 1.7. This regards the successfully converted data. The small amount of entries that was consciously rejected (contained within `wontfix` archive) has not been updated to 1.7.

The goal of the conversion was to obtain free morphological data in the tagset of the IPI PAN Corpus of Polish (IPIC or KIPI tagset) [KIPI]. This goal was achieved in two steps: we converted the data into an intermediate tagset called *IKIPI* (intermediate KIPI tagset), created within another project[1]. Then conversion routines developed along with IKIPI were used to provide compatibility. Technically, the tagset definitions and the conversion routines (two-direction conversion between KIPI and IKIPI) are now bundled with the Maca system.

Conversion methodology

The input data were divided into parts, corresponding to grammatical class as specified in the Morfologik dictionary. Each part was split into groups according to the number of different attribute values provided. Each of such groups were treated separately. The whole conversion was carried out with the help of Maca system and standard GNU/Linux utilities (grep, sed, awk, vim). Some groups were completely revised manually by a linguist.

Duplicated entries appear in the input dictionary. These were filter by `tabclean.py` script provided with Maca.

Attribute-specific issues

Underspecified masculine gender (`m`) was treated as referring to all three possible masculine genders (`m1.m2.m3`).

The distinction between neuter gender types was discarded as it is made in [KIPI]. Similarly, *plurale tantum* genders were cast to closest corresponding gender values (`p1` → `m1` , `p2` → `n` , `p3` → `n`). The occasional `pltant` number value was cast onto `pl`.

`refl` is described as an attribute value in the Morfologik README file, although, tagset-wise, it is a class in the input dictionary. Fortunately this distinction is not preserved in the destination tagset.

Class-specific conversion procedures

Here we describe particular groups. The names consist of grammatical class mnemonic and the number of attribute values provided. The entries marked with asterisk were subjected to unsafe restoration of gender or aspect values (i.e. not enough data were provided in the input so we provided all the possible values of the missing attribute — therefore redundant tags are likely to happen in our output). The groups marked `wontfix` were consciously rejected when a whole group was decided to be too unreliable to be converted semi-automatically.

`subst-1`: forms labelled `irreg`, unable to treat them automatically (`wontfix`)

subst-2: one form with lacking number value, manually corrected.

subst-3*:

1. Most forms already compliant with IKIPI
2. pltant → pl
3. `m` → `m1.m2.m3`
4. n.n2 → n
5. Some errors remain: many negated gerunds described as `subst`, e.g. `nieaklamowanie/subst:sg:acc.nom.voc:n`.

`subst-4`*:

1 Innovative Economy Programme project POIG.01.01.02-14-013/09.

1. Depreciative forms (`subst:...:depr`), affirmative gerunds (labelled `subst:ger`), negated gerunds (labelled `subst:...:neg`) and one completely erroneous entry.
2. Semicolon notation error `acc:gen` → `acc.gen`
3. Depreciatives: `m[123]?:depr` → `m2`, `subst` → `depr`
4. Gerunds lacking aspect value. We restore the aspect automatically exploiting infinitives in the input that are fortunately specified for aspect (the same for later verb groups).
5. Negation is restored — the forms starting with *nie-* whose lemma is not starting with *nie-* are treated as negated.
6. Again `m` → `m1.m2.m3`.

`adv`

1. Forms labelled just `adv` (no values) are relabelled as `qub`. These entries are removed from `adv`.
2. Note: this is not so simple. Several forms with actual `comp` degree are also labelled as just `adv`. This was later removed from qub and restored to proper tag.

Similarly in `adv:pos:neg`.

1. Lemmas of `adv:pos`: we always take the orthographic forms as lemmas, some lemmas were erroneous (`awk '{print $1"\t"$1"\t"$3}'`)
2. We are not interested in negation of adverbs, so we discard its value.

`verb-1`*:

1. These are `imps`.
2. Restoring aspect based on lemmas (see gerunds).

`verb-2`: These are `inf`, aspect given.

verb-4:

1. These are `bedzie`, `impt`, `fin`.
2. `impt` are fully specified.
3. every `bedzie` entry is suffixed with `:fut` (this is IKIPI).
4. fin:...:perf are suffixed with :fut, `fin:...:imperf` are suffixed with `:prs`.

`verb-5`*:

1. These are labelled `verb:praet` (`praet` and `fin` from IKIPI) or `verb:winien`.
2. Again `m` → `m1.m2.m3` .
3. After changing segmentation strategy of the winien class in IKIPI (19.11.2010), two situations were considered: winien without aglt (unspecified for person) and winien + aglt (winien:pri or winien:sec taken as it is in Morfologik). Note: several entries were mislabelled for person, this was manually corrected.
4. `praet:...:ter`: got rid of `ter` and taken as `praet`.
5. Remaining `praet` relabelled as `fin` (the new IKIPI `fin`, such cases were specified for gender).

`verb-6`*:

1. These are conjunctives (`conjt` in IKIPI).
2. Again `m` → `m1.m2.m3` .
3. Substituting `verb:pot:praet:` → `conjt:` .

`adj-0`: a couple of difficult cases (wontfix)

`adj-3`:

1. We add the mussing `:pos` .
2. Again `m` → `m1.m2.m3` .

`adj-4`*: `m` → `m1.m2.m3`.

`adj-5`*:

1. Getting rid of negation values.
2. Again `m` → `m1.m2.m3` .

`adjp`: already compliant.

`pant-1`: already compliant.

`pcon-1`: already compliant.

`pact-3`*:

1. All in fact affirmative but this was not marked (adding `:aff`).

2. `m` → `m1.m2.m3`.
3. No aspect given but `pact` is inherently `imperf` (adding).

`pact-4*`:

1. No aspect given but `pact` is inherently `imperf` (adding).
2. `m` → `m1.m2.m3`.

`ppas-3*`:

1. No aspect given, restoring by lemmas (see gerunds).
2. No negation given, restoring by lemmas: if the form starts with *nie-* and the lemma does not, we mark the form as `:neg`.
3. `m` → `m1.m2.m3`.

`pred-0`: already compliant; some end in *-ż* and *perhaps* should be split (although not split in the IPI PAN Corpus).

`siebie-1`: already compliant.

`prep-0`: these correspond to `prep`+`aglt`, no way to deal with it now (`wontfix`).

`prep-1`:

1. Several prep+aglt (`wontfix`).
2. Vocalicity not given. We manually revise the entries where lemma is different than form. The remaining entries are corrected automatically: if the same lemma appears elsewhere with `:wok`, we suffix the form with `:nwok`.

`qub-0`:

1. Loads of nouns and other perfectly declinable forms that must have made their way into qub accidentaly.
2. Heuristic rule: manually revising only the forms whose lemmas coincide with forms.
3. One of the most frequent qub is clearly missing: *się*. Adding it manually (with the same lemma).

`ign`: mostly abbreviations, no systematic way to handle them (`wontfix`).

`nstd`: single non-standard form *domie* (`wontfix`).

Future work

Three classes were theoretically available in Morfologik but we left them unconverted (the reasons described above): refl, ign, nstd.

The following classes were converted only partially:

1. `adj-0` (indeclinable adjectives) ,
2. some `conj` entries (`conj`+`aglt`) ,
3. `prep-0` (prep+aglt) ,
4. `subst-1` (claimed `irreg`, declinable forms in fact) ,
5. qub-0 (many inflected forms).

The following classes exist in the IKIPI and KIPI tagsets but are not represented in the converted dictionary:

1. `adja` (could be acquired semi-automatically from unannotated corpora),
2. `aglt` (few forms, could be described manually),
3. `interp` (not needed, Toki/Maca recognises punctuation)
4. `numcol` (present among `num`)
5. `xxs` and `xxx` (some candidates are labelled `qub`) .

Acknowledgements

The project is financed by the National Centre for Research and Development (NCBiR) agreement SP/I/1/77065/10.

We would like to thank Marcin Miłkowski for his help in the described work.

References

[MACA] Adam Radziszewski and Tomasz Śniatowski, *Maca: a configurable tool to integrate Polish morphological data*, FreeRBMT11, Barcelona, 2011. http://hdl.handle.net/10609/5645

[KIPI] Adam Przepiórkowski, The IPI PAN Corpus: Preliminary version, Warsaw, 2004 . http://nlp.ipipan.waw.pl/~adamp/Papers/2004-corpus/