Implementing an Old Czech Word Forms Generator

Ondřej Svoboda

Czech Language Institute of the Czech Academy of Sciences
Valentinská 91/1, 116 46 Praha 1, Czech Republic
svoboda@ujc.cas.cz

Abstract. This paper presents a word forms generator for Old Czech, originally covering common nouns but extended to produce forms of other POS, with their specifics (adverbs, prepositions, verbs). After describing the background, it gives an account of the development process and dives into the steps the generator performs. Two fundamental components of the generator are then shown in detail. Apart from the current status, future steps are also suggested.

Keywords: Old Czech, word forms generator, lemmatization, implementation

1 Preface: formal description of declension of Old Czech common nouns

In 2017, the first part of formal description of Old Czech[1] inflectional morphology was conceived (including the lexicon): Pavlína Synková concluded her dissertation [4] on declension of Old Czech (OC) common nouns, created, from the very start, with algorithmic implementation in mind.

The intended use of the implementation was (and still is) automatic lemmatization of "Old Czech text bank" (from now on, called "the corpus"), an ever-growing collection of manually transcribed and edited Old Czech manuscripts and prints (the "reference material")[2] hosted at *Vokabulář webový* ("VW")[3] since 2006.

1.1 Origins of the formal description

Building on detailed description of declension patterns of common nouns given by modern grammars, Synková first used a corpus-based tool to validate, extend and formalize the description with regards to the reference material[4] and

[1] Old Czech period spans from around 1150 to 1500.

[2] In contrast to transliteration, the text is rendered in modern orthography and error-corrected while preserving features of OC, such as the phoneme *ě* (yat) in *cěsta* ("path, road, journey"; Modern Czech: *cesta*).

[3] https://vokabular.ujc.cas.cz

[4] She mostly worked with a non-public, bigger version of the corpus (containing "alpha quality" transcriptions) but consulted the manuscripts many times (in doubt or when handling rare declension patterns), improving the corpus' quality in the process.

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2019, pp. 15–21, 2019.

second, she carefully assigned the formalized declension patterns (from now on, "paradigms") to almost 30,000 common noun headwords extracted from modern dictionaries also hosted at VW.

To handle variation in *inflectional bases* she introduced a few kinds of "alternation", assigned to individual headwords. E.g., in an E—∅ alternation, when comparing *o**kn**-o* ("window") and *o**ken*** (in genitive plural), "kn" (symbolically, `KK`) alternates with "ken" (`KeK`; `K` stands for a consonant).

She covered language change by developing a set of targeted search & replace rules ("sound changes") to transform early OC surface forms into late OC forms. To post-process raw productions into (orthographically) correct forms, she designed further rules ("formal changes"), allowing the paradigms and headwords to contain only the linguistically relevant, compact information.

E.g., *ušiech* (a local plural form of *ucho*, "ear") is produced from **uch-iech*, composed of an inflectional base and a regular ending, by turning "chie" into "šie". Afterwards, the form develops into *uších* ("ie" > "í"). Using the two rules avoids, respectively, the need for, possibly, a special case in alternation (dual forms use an *uš-* base; including also certain plural forms would be complicated), and also the need for a more modern ending (*-ích*).

2 Continuous extension & implementation

Since later in 2017, work has been underway on both sides: a word forms generator was written in .NET/C# to build a morphological database, eventually used to annotate the corpus with a home-grown tagger (in Python), handling specifics of the corpus[5].

To drive initial development and contribute to quality and stability, a list of more than 100 headwords (at least one for each paradigm) was compiled to check for correct as well as defective productions, helping to spot regressions.

A simple web interface[6] was created to provide interactive access to the generator, inspect word forms (arranged in tables) and their properties, and discover errors not yet covered by the regression test.

After "proving the concept" with a working implementation covering common nouns, further headwords and paradigms have been added, prompting changes at the generator's side. Additions so far include all immutable POS (prepositions, conjunctions, particles, and interjections), adverbs (both gradable and ungradable), and a few large conjugation classes of verbs.

2.1 Features of headwords

With nouns, alternations of the inflectional base were introduced, as well as a simple paradigm constraint (singular- or plural-only). Adverbs can declare

[5] Such as using occasional manual lemmatization by editors of the source manuscripts to partially disambiguate otherwise highly ambiguous automatic annotation. (No OC training corpus is currently available to me although one existed in the past [1].)

[6] `https://ridics.ujc.cas.cz/nlp/word-forms/`

(often suppletive) gradated forms in place of a regular paradigm. Similarly, prepositions list the governed cases and vocalized ones specify vowels to attach (*k*, *ke*, *ku*).

Verbs provide their aspect and will require multiple types of alternation (e.g. vowel quantity and modification of the final consonant cluster) to apply simultaneously.

3 From a headword to word forms

This section describes the process of generating word forms from a headword, given its POS or paradigm ID, and other properties, mentioned above, as input.

3.1 Paradigm setup

After the paradigm is either looked up by its ID, constructed at runtime (for prepositions and irregular adverbs), or an invariant one is used, it is accomodated to the headword and its constraints. For example, *húsle* ("fiddle") selects only plural endings of an otherwise complete paradigm *kost*.

Terminology: Since endings of some paradigms also come with suffixes, a more general concept will be used in the following text. "Terminations" *proper* refer to endings with suffixes. Generalized "terminations" can, in addition, carry further morphs such as the superlative prefix (*naj-*), various forms of the negative prefix, and even suppletive inflectional bases (in verbs like *býti*, "to be").

Sound changes are applied to the paradigm, creating (proper) terminations more recent than the initial, early 14th century ones. The changes will affect the additional morphs specified above, in the near future.

3.2 "Stemming"

Before an inflectional base can be retrieved, special, "stemming"-only terminations are also created when performing sound changes, to account for a pre-1300 *lě* > *le* sound change, and related ones. For *húsle*, the closest suitable termination in `PL.NOM` is *-ě*, which cannot be removed directly. For this purpose, an *-e* termination is developed, requiring "l" before it. The "sound change" used here is written with regex notation: `(?<=l)ě > e`.

After the (longest) stemming termination is removed, the resulting raw base undergoes a round of supplementary formal changes to suit the paradigm. These affect headwords like *mládě* ("a young [being]"). After losing the *-ě* the base-final consonant is softened (> `mláď`) in order to produce e.g. a plural form *mláďata*.

3.3 Alternations

At this stage, alternations of the inflectional base are handled, distributing terminations of the headword's paradigm between resulting multiple variant bases. For *okno*, the ending *-o* is first removed from the initial base "okn".

The headword's E—∅ alternation is specified as `E0- zero_endings-KeK / nonzero_endings-KK`. First, the *-o* ending is matched against the `nonzero_endings` selector, with the respective `KK` pattern. Second, the two placeholders resolve to "k" and "n", and the remainder of the base ("o", an immutable "prefix") is also stored.

By applying the obtained consonants to the `KK` and `KeK` templates and appending the results to the stored "prefix", two variant bases (*okn-* and *oken-*) form. Finally, all non-zero endings are associated to the *okn-* base, while the *oken-* base is joined by a *-*∅ `PL.GEN` ending.

3.4 Further base changes

Sound changes are applied to the inflectional bases next. In addition, a *ne-* prefix is added for "negatable POS" (currently, verbs). For verbs like *obalovati*, this results in a chain of *obal-* > *vobal-* > *nevobal-* bases.

3.5 Emission

For each base, early OC word forms are created first by joining prefixes with the base and suffixes (all of various origin). Afterwards, sound changes apply to the boundary between the base and suffixes (both proper ones like *mořě* ("sea") > *moře*, and auxiliary ones like **húslě* > *húsle*), and subsequently to the whole word forms: *púšče* ("deserts") > *púště*.

Throughout the whole process, extra information is attached to terminations, inflectional bases, and the resulting word forms, such as sound changes applied to the respective units. Each word form is thus aware of its "origin", allowing to produce a tag made up of common grammatical categories. In the future, an additional tag distinguishing the word form from others with an identical grammatical tag (following ideas of [3]) should be possible.

4 Components of the generator

Adjusting to the need to generate word forms of various POS, the initial noun-specific implementation has grown a few interesting features and components.

4.1 Paradigms module

The generator is capable of modeling paradigms of any Old Czech POS in a single framework, while catering to their needs. Starting with a fixed two-level structure sufficient to model paradigms of nouns (three numbers with

seven cases each), the framework was generalized to house case-governing prepositions, gradable adverbs (with a capability to define regular creation of superlatives from comparatives), and highly inflected verbs.

An example below shows parts of a paradigm of verbs conjugating like *pračĕvati*. Ancestors of `<termination/>` nodes describe the structure (of uneven depth) common to all Old Czech verbs, defined in a separate meta file.

```
<present type="PRES/I">
  <singular number="SG/S">
    <firstPerson person="1/1">
    ...
<supine type="SUP/U" matchAspect="IPFV"/>
<participle type="PART">
  <nt subtype="NT" likeParadigm="declension" type="/S"
      substrate="verb.6.kupovati"/>
  ...
  <l subtype="L" type="/A">
    <singular number="SG/S">
      <termination gender="M">
        <stemSuffix>ěva</stemSuffix>
        <participleSuffix>l</participleSuffix>
        <ending>0</ending>
      </termination>
```

Some of the structure nodes' attributes carry two values. In `@type`, `@subtype`, `@number`, and `@person`, the first part is a "glossing abbreviation"[7] instrumental in referring to a paradigm node by a path. For instance, `PRES.SG.1` is used in verbs such as *prositi* to change the root-final consonant from "s" to "š", producing *prošu*, and `PL` is used to select a part of a paradigm in plural-only nouns. Following the slash, the second part is a value of the Czech attributive tagset [2], giving (partial) tags `nSp1mI`, `mU`, and `gMnSmA` for the above example.

The `@substrate` attributes allow (nominal and verbal) paradigms to inherit terminations from (parts of) other paradigms. The `@matchAspect` attribute guards against generating supines for perfective verbs.

The `<termination/>` nodes are found in leaf nodes of the structure and contain a complete set of morphs to be attached to (or even replace) an inflectional base, already introduced in 3.1. This allows to generate *nenie*, *nejnie*, and *nénie* (`PRES.SG.1.NEG`) for *býti* or use full, supplied forms for complex cases like *týden* ("week"): *téhodne*.

4.2 Sound changes engine

This component has been shown to be essential to several stages of word forms generation. So far, sound changes have targeted terminations ("proper" only,

[7] `https://en.wikipedia.org/wiki/List_of_glossing_abbreviations`

not "generalized"), inflectional bases (both "raw" and post-alternation ones), the boundary between the base and suffixes, and whole word forms.

Up to now, about 50 out of 100 changes defined in the dissertation have been enabled in the generator.

As an example, the `u-i-change`, representing a pair of sound changes written as *'u* > *'i* and *'ú* > *'í*, is accompanied by two kinds of information, relevant to the generator, or only to users. It lists its targets (inflectional bases, base—suffix boundaries, terminations), a contemporary/related change (`o-yat-change`), the approximate time span it was in effect (2nd to 3rd fourth of the 14th century), and finally its specification (here in a redacted form), using a regex look-ahead to block it from applying on the *uo* digraph:

`Ku(?!o) > Ki` and `Kú > Kí` where

`K = ž,š,č,ř,j,ď,ť,ň,ḃ,ḟ,ľ,ṁ,ṗ,ṡ,v́,ź,c`

It could also define a short value to use in a "variant/mutation" [3] tag to distinguish the input form (*zeḿu*) from an (intermediate) output (*zeḿi*).

The lower bound of the time information could be used in two ways: to attach an estimate datation ("not before") to the targets, and enforce a chronological order when applying sound changes.

5 Conclusion

This article presented a generator capable of producing Old Czech word forms on the grounds of a list of headwords (with POS-specific properties), a linguistically adequate repertoire of inflectional paradigms (as compact as possible), and a set of sound changes targeted at concrete morphemes.

During the process, the generator keeps track of the origins of each word form, enabling it to attach useful, rich information to its output.

Acknowledgements This work was fully supported by RIDICS ("Research Infrastructure for Diachronic Czech Studies"), a project LM2015081 funded by the Czech Ministry of Education, Youth and Sports of the Czech Republic.

References

1. Hana, J., Lehečka, B., Feldman, A., Černá, A., Oliva, K.: Building a corpus of old czech. In: Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC. p. 1–7. European Language Resources Association, Istanbul (2012), `https://msuweb.montclair.edu/~feldmana/publications/2012-morph-lrec.pdf`
2. Jakubíček, M., Kovář, V., Šmerk, P.: Czech morphological tagset revisited. In: Proceedings of Recent Advances in Slavonic Natural Language Processing. p. 29–42. Tribun EU, Brno (2011), `https://nlp.fi.muni.cz/raslan/raslan11.pdf#page=37`
3. Osolsobě, K., Hlaváčová, J., Petkevič, V., Šimandl, J., Svášek, M.: Nová automatická morfologická analýza češtiny [the new automatic morphological analysis of czech]. Naše řeč, AV ČR, Ústav pro jazyk český **100**(2), 225–234 (2017)

4. Synková, P.: Popis staročeské apelativní deklinace (se zřetelem k automatické morfologické analýze textů Staročeské textové banky) [Description of Old Czech Common Nouns Declension (with regard to Automatic Morphological Analysis of Texts in Old Czech Text Bank)]. Ph.d. thesis, Charles University, Faculty of Arts, Institute of Czech Language and Theory of Communication, Praha (2017), http://hdl.handle.net/20.500.11956/86563