Petr Plecháč*

Czech Verse Processing System KVĚTA – Phonetic and Metrical Components

Abstract: The following paper describes the algorithms of phonetic and metrical components of the Czech verse processing system KVĚTA, updating information contained in previous reports (Ibrahim and Plecháč 2011; Plecháč et al. 2013a; Ibrahim and Plecháč 2014). The system is being used in the building of the Corpus of Czech Verse (hereinafter CCV), which at present contains 1 689 Czech books of poetry (over 2.5 million lines) from the nineteenth and early twentieth centuries. In contrast to standard language corpora, in each lexical unit are not only the lemma and morphological tag attributes assigned but they also contain a phonetic transcription; furthermore, the attributes metre (iamb, trochee...), length (number of feet), ending (feminine, masculine...) and metrical pattern are assigned to each verse line. At higher levels rhyme pairs (or n-some) and fixed forms (sonnet, rondel, etc.) are annotated. Here we will focus on components providing phonetic and metrical annotation: (1) the F-component, whose task is to derive the phonetic transcription from the input data, (2) the G-component, whose task is to generate a set of all possible metrical interpretations of these data, and (3) the M-component, whose task is to select from this set the final interpretation. Automatic analysis has so far been limited to accentual-syllabic (hereinafter AS) and monometric poems – i.e., poems consisting of repetitions and variations of a single metrical pattern (though AS imitations of some quantitative meters are recognized).

Keywords: Verse Processing, KVĚTA, Czech language, phonetic and metrical annotation

1 Input data

On the basis of metadata and/or graphical elements, KVĚTA first divides the input data into stanzas, lines and single words (tokens). Lemmatized and

*Corresponding author: Petr Plecháč, Institute of Czech Literature, Czech Academy of Sciences, Na Florenci 3/1420, 110 00 Prague, Czech Republic, E-mail: plechac@ucl.cas.cz
morphologically annotated data are required, i.e., each token needs to be provided with its *lemma* and *morphological tag* (in the format Prague positional tagset format). Of the latter only two positions are used in three components discussed at present – position 1: Part of speech (D: adverb, N: noun, V: verb...) and position 11: Negation (A: affirmative, N: negated, -: not applicable). For example the *ad hoc* created line “Kdy chrti cení zuby” is therefore provided as:

<table>
<thead>
<tr>
<th>token</th>
<th>lemma</th>
<th>morphological tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kdy</td>
<td>kdy</td>
<td>Db---------------</td>
</tr>
<tr>
<td>chrti</td>
<td>chrt</td>
<td>NNMP1------------</td>
</tr>
<tr>
<td>cení</td>
<td>cenit</td>
<td>VB-P---------------</td>
</tr>
<tr>
<td>zuby</td>
<td>zub</td>
<td>NNIP4------------</td>
</tr>
</tbody>
</table>

(1) The example of input data

2 F-component

The core of the F-component, which provides the phonetic transcription, consists of a sequence of commands that successively overwrite the input data. These commands are based on well-known and well-defined rules, such as assimilation of voicing, palatal pronunciation of bigrams *di, ti, ni* etc. For example, when transcribing the above-mentioned line, the following commands are used:

---

1 Lemmatization and morphological annotation were carried out by researchers at the Institute of Theoretical and Computational Linguistics, Faculty of Arts, Charles University in Prague (Hana Skoumalová, Milena Hnátková, Tomáš Jelinek and Vladimír Petkevič) in cooperation with researchers at the Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague (Jan Hajič, Jaroslava Hlaváčová).


3 For a detailed list of rules on which the algorithms of the F-component are based, see Palková 1994: 320–345.

4 SAMPA is used for phonetic transcription throughout the paper (see <http://noel.feld.cvut.cz/sampa/>).

Beside these phenomena which can be captured by a finite number of rules, there are also cases in which a taxative approach has to date seemed more reliable. This includes the transcription of the bigrams au, ou, eu and the transcription of loanwords.

### 2.1 Bigrams au, ou, eu

Unless individual morphemes are annotated, there seems to be no efficient way in Czech of automatically deciding which instances of the bigrams au, ou, and eu represent a diphthong (pouzdro, kauza, eukalyptus) and which represent two stand-alone vowels (pouliční, ponaučení, přeudatný). All these instances are therefore transcribed right after the upper case → lower case command by means of a manually built library\(^5\) with a simple structure token → bigram(s) transcription:

<table>
<thead>
<tr>
<th>Input</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>pouzdro</td>
<td>o_u</td>
</tr>
<tr>
<td>pouliční</td>
<td>ou</td>
</tr>
<tr>
<td>kauza</td>
<td>a_u</td>
</tr>
</tbody>
</table>

The exceptions are as follows:

1. The bigram eu whose first letter belongs to the negative prefix ne- (neurčitý, neukázal). These cases are distinguished from other tokens beginning with the trigram neu (neurastenický) by position 11 of morphological tag (see Hajič 2004: 82) and thus may be automatically transcribed as [eu].

---

\(^5\) The library currently contains almost 40,000 tokens and is being further extended.
(2) The bigrams *au*, *ou*, *eu* (a) followed by grapheme *j*, (b) preceded by a sequence of more than three non-syllabic consonants, or (c) preceded by a sequence of three non-syllabic consonants the last of which is either [j] or a palatal consonant. In such contexts these bigrams cannot represent diphthongs (Bičan 2013: 298) and may thus automatically be transcribed as [au], [ou], [eu].

(3) The word-final bigram *ou*, which can be reliably transcribed as a diphthong. (In the CCV there are more than 15,000 different tokens (types) ending with *ou*. According to a manual control of all the occurrences, this bigram represents two stand-alone vowels in this position only in three rather peculiar instances: *Athou* – loc. sg. of the lemma *Athos* (perhaps typo – missing *s*: *Athosu*), *Elou* – probably acc. sg. of the name *Eloisa*, and *herou* – dat. sg. *heroj*).

There are also several forms of poetic license that it is impossible to include in the library. This concerns for example the occasional disyllabic treatment of *nauka* ([na_uka] instead of the common [nauka]),⁷ the monosyllabic treatment of *Zeus* ([ze_us] instead of the common [zeus]), etc. A similar case is the treatment of preterite verbs such as *nesl*, *vedl* as apocopations ([nes] instead of [nesl = ], [vet] instead of [vedl = ]) which is very common at the turn of the nineteenth century. Such unsystematic cases are usually revealed within a manual control initiated by the M-component (see Section 5.2).

### 2.2 Loanwords

In dealing with loanwords it would be theoretically possible to consider some combined stochastic/rule-based algorithm that would find the most probable donor language based on occurrence of a specific grapheme (č, ž, etc.) or n-gram (ais, aet, etc.) and then apply the transcription rules of this language instead of those of Czech. There are, however, pragmatic counter-arguments:

---

⁶ One could object that it would be also possible to transcribe all the word-initial bigrams *au*, *ou*, *eu* as diphthongs. Their occurrence in the beginning of the word is, however, rare in the Czech language and in many cases it indicates a loanword (see Section 3.2). Moreover, in older texts the initial *o* may represent the graphically non-separated preposition (*oumlvéh, oudatné*), i.e., a stand-alone syllable.

⁷ This and similar following statements are (of course) based on the expectancy of isosyllabicity of lines.
(1) Loanwords represent only a negligible part of the analysed material.
(2) Although the number of possible donor languages is theoretically finite it is rather large.
(3) Among the most frequent donor languages there are those with highly non-phonemic orthography (French, English).
(4) Many loanwords are partly adapted to Czech orthography (see Section 3.3), so using either the rules of the donor language or rules of Czech would not automatically produce a correct output.
(5) The error rate of the stochastic part of the algorithm would probably be rather high.

Loanwords are therefore also transcribed by means of a manually built library. They are detected as already mentioned – by the occurrence of graphemes not contained in the standard Czech alphabet, occurrence of the unusual n-gram – or by metrical irregularities caused by incorrect transcription (see Section 5.2).

Yet another problem with loanwords is that they usually exhibit a high degree of homography (e. g., nom. sg. Shakespeare pronounced in Czech as [Sejkspi:r], while voc. sg. Shakespeare as [Sejkspi:re]) and even the different treatment of the same word (e. g., nom. sg. Baudelaire occurs both as di- and trisyllabic). The selection from these variants is also done manually in individual cases.

2.3 Accuracy

Basing on the manual control of a random sample of 3,000 lines containing 71,257 sounds, we estimate the accuracy of the phonetic level of annotation of CCV as being as high as 99.91% (i. e., approximately 9 out of 10,000 sounds are incorrect). The most frequent type of error (over 40% of all those found) is the incorrect palatal transcription of the bigrams di, ti, ni in non-recognized loanwords, e. g.:

\[
\begin{align*}
\text{melodicky} & \rightarrow *[\text{melo}]\text{icki} & \quad // \text{correct: [melodicki]} \\
\text{kantilény} & \rightarrow *[\text{kancile}]\text{ni} & \quad // \text{correct: [kantile:ni]} \\
\text{definice} & \rightarrow *[\text{defijit}]\text{se} & \quad // \text{correct: [definit_se]}
\end{align*}
\]

Let us mention that none of the errors found causes a miscount of number of syllables.
3 G-component

The goal of the metrical analysis provided by the G-component and the M-component is to reveal the regular organisation of certain features carried by individual syllables. Such regularities may be of four types:

1. certain alternations of strong (S) and weak (w) positions that are repeated in every line (standard AS metres),
2. certain alternations of S and w that are repeated in a group of lines (AS imitations of quantitative syllabic metres; see Gasparov 1996),
3. certain 6-tuples of \{S, Sw, Sww\} that are repeated in every line (AS imitations of quantitative dactylic hexametre and pentametre),
4. certain 12-tuples of \{S, Sw, Sww\} that are repeated in every two lines (AS imitations of quantitative elegiac couplet).

The task of the G-component is to extract the individual syllables from the phonetic transcription provided by the F-component, to assign these syllables attributes important to the metre and to generate all the metrical patterns to which the syllables may correspond. Since the metrical organisation is usually far from being deterministic, the patterns are generated regardless of given attributes based just on the number of syllables (so no possible pattern is excluded). The final choice of most likely pattern is then done by the M-component.

3.1 Syllable classes

In AS versification the most important feature is by definition the potential prominence of syllable (stress). Usually each initial syllable of a polysyllabic word in Czech is considered to be stressed while each non-initial to be unstressed. As for monosyllabic words, the prominence depends on various factors. The most general tendency is for the content word to be stressed, while the function words are unstressed (Palková 1994: 282). A special case is monosyllabic prepositions proper (MPP), which usually behave as forming a single word with the following one, i.e., in being stressed distract the potential stress of the following syllable (\emph{na slabice}: [“naslabit se”]). This tendency is, however, not constant in Czech. Sometimes the MPPs are realized as a standard function word (i.e., an unstressed one with no effect on the following syllable; see Skarnitzl 2014) and native speakers do not perceive it as mistake. This has been broadly exploited by poets, particularly in the second half of nineteenth century (see Plecháč et al. 2013b). It has also been shown that when a potentially prominent syllable occurs in the position where one would expect a non-
prominent one, in many texts, mainly from the first half of the nineteenth century, the syllable containing the long vowel is statistically significantly preferred in the following position (Plecháč et al. 2013b), e. g.:

Mládež za mrazu šedává
S w S w S w S w

Finally the same case is usually considered less irregular when a monosyllabic word occurs in the position preceding the potentially prominent syllable, e. g.:

Aj! kdo zná ji, tu osobu
S w S w S w S w

G-components thus first create a two-dimensional array of objects, each of which corresponds to one syllable in a phonetic transcription, or more precisely to one sound from the set

{a, e, i, o, u; a:, e:, i:, o:, u:, a_u, e_u, i_u, o_u, e_u, r =, l =, m =, n =, s =, S =, z =}

For every poem that has m lines each of which has n[m] syllables the following data structure is created:

```
syll = [
    [syl1[1][2], ... syl1[n[1]]],
    [syl2[2][2], ... syl2[n[2]]],
    ...
    [sylm[2][2], ... sylm[n[m]]]
]
```

The following Boolean attributes are then assigned to each object:

1. initial // 1: word-initial syllable, 0: non-initial
2. final // 1: word-final syllable, 0: non-final
3. contentWord // 1: content-word, 0: function word
4. preposition // 1: MPP, 0: other

8 For the sake of clarity we have used one-based indexing here instead of more common zero-based indexing, i. e., the first element of an array is indexed 1 instead of 0.
9 The following parts of speech are counted as content words: noun, adjectives, numerals, interjections, verbs (except for forms of the lemma být).
10 The following tokens are counted as MPP if disambiguated as prepositions (position 1 of morphological tag): před, od, ob, ku, ke, do, ve, po, nad, přes, při, bez, se, ze, za, u, pod, pro, zpod. In the case of two subsequent MPPs, the first one is counted as a standard function word. A longer sequence of MPPs is not permitted by the morphological tagger (see Jelinek and Petkevič 2011: 162).
(5) prevPreposition   // 1: preceded by MPP, 0: preceded by other
(6) prevInitial      // 1: preceded by word-initial syllable, 0: preceded by other
(7) nextLong         // 1: followed by syllable containing long vowel, 0: followed by other

For example, the attributes of the third syllable of the aforementioned line Ají kdo zná ji, tu osobu would have the following values:

```plaintext
syll[1][3] = {
    initial: 1,
    final: 1,
    contentWord: 1,
    preposition: 0,
    prevPreposition: 0,
    prevInitial: 1,
    nextLong: 0
}
```

Theoretically there are $2^7 = 128$ different combinations of attributes values. Some combinations are, however, excluded. This concerns e. g. *{initial: 0, preposition: 1} or *{final: 0, preposition: 1} since a syllable constituting a monosyllabic preposition proper is necessarily both word-initial and word-final. In many cases some attributes are not important at all (see above). For example, the attribute contentWord is used to classify monosyllabic words and in other cases it is of no importance. Furthermore, if the monosyllabic word is preceded by MPP {prevPreposition: 1}, the attribute contentWord is also of no importance. The attributes nextLong and prevInitial are taken into account only if {initial: 1, final: 0} or {preposition: 1}. Many more such combinations are grouped into one class and as a result only 12 classes (syllClass) are recognized.

### 3.2 Standard AS metres

For every poem except for those described in Section 4.5 the patterns of all standard AS metres are generated. The algorithm is as follows: let max(syll) be the number of elements in the largest syll[1..m] (i. e., the longest line). Then the list of all possible strings having max(syll) characters and matching the regular expression

```
/^w?(Sww?)*(Sw?)?$/
```

---

11 Diphthongs are counted as long vowels.
is generated, i.e., all possible combinations of iambic foot (wS), trochaic foot (Sw), dactylic foot (Sww), amphibrachic foot (wSw) and line-final S-position. All these strings are used as keys in the associative array metre.

Next a two dimensional array of metrical positions, each of which corresponds to one syllable, is created for each metre on the basis of initial substring of the given key. Thus e.g., for a quatrain consisting of 6-syllable line – 3-syllable line – 6-syllable line – 3-syllable line (max(syll) = 6) the following metres are generated:

‘SwSwSw’, ‘SwSwS’, ‘SwSww’, ‘wSwSwS’, ‘wSwSww’, ‘wSwwSw’

For the first metre following array is created:

```java
metre[SwSwSw] = [
    ['S', 'w', 'S', 'w', 'S', 'w'],
    ['S', 'w', 'S'],
    ['S', 'w', 'S', 'w', 'S', 'w'],
    ['S', 'w', 'S']
]
```

### 3.3 AS imitations of quantitative syllabic metres

AS imitations of quantitative syllabic metres are based on repeating groups of lines in which each n-th line exhibits the same syllabic length and the same metrical pattern. So far the four most common metres are recognized by KVĚTA – Sapphic stanza, Asclepiad III, and two types of Alcaic strophe. If the poem fulfils syllabic requirements, the following arrays are created in the G-component:

```java
metre[sapphic] = [
    ['S', 'w', 'S', 'w', 'S', 'w', 'S', 'w', 'S', 'w'],
    ['S', 'w', 'S', 'w', 'S', 'w', 'S', 'w', 'S', 'w'],
    ['S', 'w', 'S', 'w', 'S', 'w', 'S', 'w', 'S', 'w'],
    ['S', 'w', 'S', 'w', 'S', 'w'],
]
metre[asclepiad] = [
    ['S', 'w', 'S', 'w', 'S', 'w', 'S', 'w', 'S', 'w', 'S', 'w', 'S', 'w', 'S', 'w'],
    ['S', 'w', 'S', 'w', 'S', 'w', 'S', 'w', 'S', 'w'],
    ['S', 'w', 'S', 'w', 'S', 'w'],
]
metre[alcaicA] = [
    ['w', 'S', 'w', 'S', 'w'],
]```
3.4 AS imitations of quantitative dactylic hexametre, pentametre and elegiac couplet

The situation with AS imitations of quantitative dactylic hexametre and pentametre is more complicated. In a hexametrical poem by definition every line exhibits a pattern matching the regular expression

\[^{S\text{ww}S\text{ww}S\text{ww}S\text{ww}S\text{ww}S\text{ww}S\text{ww}S}\].\(^{12}\)

The number of syllables may, therefore, vary from 12 (SwSwSwSwSwSwSw) to 17 (SwwSwwSwwSwwSwwSwSw). While in all the cases discussed so far the metrical pattern has been well-defined by the type of metre and the number of syllables in line (n) in hexametre this clearly holds true only for \(n \in \{12, 17\}\). For \(n \in \{13, 16\}\) there are 5 different patterns that match the abovementioned regular expression, for \(n \in \{14, 15\}\) there are 10 such patterns. Therefore, if the number of syllables in all the lines of the poem vary from 12 to 17 (e. g., it is suspected to be hexametrical), the array \text{metre}[\hexametre]\ is defined as three-dimensional. Thus e. g., for a couplet consisting of a 13-syllable line and a 12-syllable line the following array is created:

\[
\text{metre}[\hexametre] = [
[S', 'w', 'w', 'S', 'w', 'S', 'w', 'S', 'w', 'S', 'w', 'w'],
[S', 'w', 'S', 'w', 'S', 'w', 'S', 'w', 'S', 'w', 'S', 'w'],
[S', 'w', 'S', 'w', 'S', 'w', 'S', 'w', 'S', 'w', 'S', 'w'],
[S', 'w', 'S', 'w', 'S', 'w', 'S', 'w', 'S', 'w', 'S', 'w'],
[S', 'w', 'S', 'w', 'S', 'w', 'S', 'w', 'S', 'w', 'S', 'w']
]\]

\(^{12}\) Although most poetic handbooks require the penultimate foot to be dactylic (Sww), there are dozens of hexametrical lines in the CCV that have trochee (Sw) in this position. Thus this constraint is not reflected in G-component.
Dactylic pentametre is treated similarly. In a pentametrical poem every line exhibits the pattern matching the regular expression


Therefore, if the number of syllables in all the lines of the poem vary from 10 to 15, the three-dimensional array metre[pentametre] is created.

Pentametre, however, is by far most frequently used in the so-called elegiac couplet – the regular alternation of hexametre and pentametre. For poems where the number of syllables in odd lines varies from 12 to 17 and in all even ones from 10 to 15 the three-dimensional array metre[elegiac] is, therefore created, in which the patterns of hexametre and pentametre alternate.

### 3.5 Radif

The last case which is treated differently are ghazals, i.e., poems in which the first and every even line contains a so-called radif (repeating word or group of words in the end of line). For example:

Na dvěře žíti člověk zaklepal. – Dále!
Tak rázným hlasem život hned ho zval dále.
A sotva vešel, člověk zaplakal.
Však nedbal osud, ale rozkázel: Dále!
A kam se hnul, vždy opodál
s tím zvukem osud za ním stál: Dále!

Radif is a stand-alone rhythmical unit and metrical patterns of ghazals, therefore, need not match standard AS patterns. The example above is written in iambic pentametre with masculine ending (wSwSwSwSwS) which is in the lines 1, 2, 4, and 6 followed by radif with pattern Sw. The complete pattern of these lines thus contains two subsequent S-positions (wSwSwSwSwSSw) which violates the constraints given in Section 4.2.

When dealing with poems in which the above-mentioned lines end with the same word or group of words, the G-components first cut off this repeating unit

13 The possibility of third S-position being followed by w-position (which does not occur in ancient hexametre) is pretty common in the Czech tradition.
from the rest of the line and apply the method described in Section 4.2 to each of these two parts separately. Two sets of patterns are created and the final list of patterns is then defined as their cartesian product.

4 M-component

The task of the M-component is to select from all the elements of metre the final interpretation based on attributes of elements of syll.

4.1 Metrical coefficient

The M-component first loops through all the elements w of metre and to each metrical position of each metre metre\[w\][i][j] it assigns a metrical coefficient that reflects the probability of a given syllable being the realization of a given position. When building the CCV the values of the metrical coefficient were set mainly on the basis of introspection (see Ibrahim and Plecháč 2011). Due to the fact that the CCV contains nowadays more than 2 million annotated AS lines, the values may be set more accurately on the basis of data already analyzed in the following way.

For the \( j \)-th syllable of the \( i \)-th line and metre \( w \), let \( \sigma_{i,j} \) be the syllable type of syllClass[i][j], \( x_{w,i,j} \) be the value of metre\[w\][i][j], \( \mu_{w,i,j} \) be the metrical coefficient of \( x_{w,i,j} \) realized by \( \sigma_{i,j} \), and \( P(\sigma_{i,j}|x_{w,i,j}) \) be the probability of position \( x_{w,i,j} \) being realized by \( \sigma_{i,j} \) estimated from its frequency in CCV. Since \( x_{w,i,j} \) is a binary variable with possible values S and w probability of which may be considered as being equal, following can be derived from Bayes’ theorem (under a naive independence assumption):

\[
P(x_{w,i,j}|\sigma_{i,j}) = \frac{P(\sigma_{i,j}|x_{w,i,j})}{P(\sigma_{i,j} = S) + P(\sigma_{i,j}|w)}
\]

Like any other linguistic corpus, the CCV is not a homogeneous set and probabilities estimated from the frequency in the whole corpus need not to be valid in individual subcorpora. Thus e.g., a long vowel in the second syllable of a polysyllabic unit is more likely to be the realization of S-position in the work of an author from the first half of the nineteenth century than of an author from the second half (see Section 4.1). The values of \( \mu \) are, therefore, set differently for individual authors – as a weighted geometric mean of \( P(x_{w,i,j}|\sigma_{i,j}) \) estimated from all the CCV and of \( P_A(x_{w,i,j}|\sigma_{i,j}) \) estimated from the subcorpus of given
author’s work, with the weight ratio 1 : 3. This, however, is applied only if the
author’s subcorpus contains enough evidence to generalize, or more precisely if
the frequencies in the subcorpus differ significantly from the frequencies in the
entire CCV, which is tested by Yates’ $\chi^2$ test ($\alpha = 0.001$):

$$\mu_{w, i, j} = \begin{cases} 
\sqrt[4]{P(x_{w, i, j} | \sigma_{i, j})} (P_A(x_{w, i, j} | \sigma_{i, j}))^3 & \text{if } \chi^2 > 10.83 \\
\frac{P(x_{w, i, j} | \sigma_{i, j})}{P_A(x_{w, i, j} | \sigma_{i, j})} & \text{otherwise} 
\end{cases}$$

4.2 Selection of metre

The selection of final interpretation is based on the metrical coefficients of
individual line patterns ($L_{w,i}$) and the overall metrical coefficient of the metre ($T_w$).

The probability of $i$-th line being the realization of metre $w$ can be computed
as the product of all $\mu$ assigned to individual positions of this metre. In order to
get $L_{w,i}$ which is comparable among patterns of different length, the value of this
product is normalised. $L_{w,i}$ is thus computed as:

$$L_{w, i} = \prod_{j=1}^{n_i} \mu_{w, i, j}^{n_j}$$

In case of hexametre, pentametre and elegiac couplet the pattern with the
highest $L_{w,i}$ is selected in this step as the most likely one:

$$L_{w, i} = \max \{L_{w, i, 1}, L_{w, i, 2}, \ldots\}$$

$T_w$ is computed as the geometric mean of individual $L_{w,i}$:

$$T_w = \sqrt[2]{\prod_{i=1}^{m} L_{w, i}}$$

As the final interpretation the metre(s) with the highest $T_w$ is selected. For
example, the following quatrain would be this way annotated as a trochee:

14 Similarly so-called substitutions are treated in standard AS metres (e. g., SwwSwSww in the
context of SwwSwSwwSwSww). Since in standard AS metres this is rather rare and non-systematic, its
occurrence lowers the value of $L_{w,i}$.

15 The selection is also governed by several additional rules, e. g., if a hexametre ranks among
the metres with the highest $T_w$ and in each line it exhibits the pattern SwSwSwSwSwSw, such
interpretation is excluded and the poem is annotated as being a standard AS trochee. Further
rules concern alexandrine, several line beginnings as well as line endings.
(3) Selection of metre

In practice automatic annotation is done only if several conditions are met. This includes the minimal required value of $T_w$, maximal allowed standard deviation of $L_{w,1}$, $L_{w,2}$, ..., minimal value required for each $L_{w,1}$ and others. If some of these conditions are not met, the M-component requires manual control. The purpose of such conditions is to minimize the incorrect annotation of polyrhythmic and other than AS poems and to reveal possible mistakes in phonetic transcription.

4.3 Accuracy

Because of these strict conditions and a lot of manual controls/corrections done when building the CCV, we may assume that the number of errors in the CCV is in this respect very low. Basing on manual control of a random sample of 300 poems, we estimate the accuracy of metrical level of annotation in the CCV to be as high as 99.97% (8 out of 25,779 lines controlled were found to be annotated incorrectly).

If we disregard these assumed minor errors, the CCV may be used to estimate the accuracy of the algorithm used by the M-component itself, i.e. we can find in how many cases the M-component annotates the line wrongly when running without the aforementioned conditions. Out of all 2,336,435 AS lines contained in the CCV, the annotation done solely by the M-component differs from the one in the CCV in 108 979 cases. Thus accuracy may be estimated as 95.34% (with the original introspectively set values of $\mu$ (see Section 5.1) the accuracy is 94.88%).
Let us add that most of these errors concern polymetric and/or multimetric poems (i.e., poems meeting the conditions of two or more metres at the same time; see Ibrahim and Plecháč 2011; Pilshchikov and Starostin 2011). The M-component is, however, not yet designed to annotate polymetric texts automatically (improvement of the component annotating rhyme pairs is necessary). As for multimetric poems, one can rather think of differences in annotation as resulting from different approach rather than errors. If we exclude polymetric poems as well as the poems recognized either in the CCV or by the M-component as being multimetric, the accuracy of the M-component increases to as high as 97.92% (43,786 out of 2,104,039 lines incorrectly annotated).

5 Conclusions

The algorithms presented here yield satisfactory accuracy, though there is still a lot to be done, especially:

1. Improving the F-component by adding further contextual rules, such as Section 3.1.(2).
2. Extending the set of possible patterns of AS imitations of quantitative syllabic metres in G-component.
3. Improving the component that annotates rhymes and extend the M-component in order to be capable of correctly processing texts written in two or more different metres.

Funding: Supported by the Czech Science Foundation (P406/11/1825) and the research institution 68378068.

References