Generating Czech Iambic Verse

Zuzana Nevěřilová and Karel Pala

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

Abstract. In the paper, we describe an algorithm for generating Czech iambic verse and its implementation on a computer. It is a continuation of the work first done in 1972 [6], in which a program generating Czech iambic verse had been developed, written in the programming language Algol-Genius and run on the mainframe SAAB D21 with interesting results.
Here, we present a new experiment which is a follow-up of the previous one and includes some new techniques as e.g. using a complete Czech automatic morphology (Majka) plus small DC grammar of Czech. The poetic rules built before 1972 by Jiří Levý are used with only slight modifications. SWI Prolog has been selected as a programming language. Also the resources are different, as a base for the poetic vocabulary we have decided to use the literary texts from Czech Wikisources.

Keywords: poetry, natural language generation, definite clause grammars, semantics

1 Introduction

The goal of the paper is to model on a computer the creative processes, in which Czech iambic verse is produced. We describe an algorithm for generating Czech iambic verses and the implementation of the whole procedure. In other words, we are interested in some general aspects of computer modeling of some human brain functions.

A computer generating of Czech verse was initiated by J. Levý in 1966 who outlined the three main objectives:

- Formulating the poetic rules which would enable us to generate Czech verses with various metrum on the computer: the first attempt was made for the ten-feet iambic verse by J. Levý shortly before he died (on January 1967, [4]).
- Building the proposed poetic rules into a formal (context-free) grammar of Czech and implementing it as a program generating Czech sentences after confrontation of grammar and poetic principles.
- Analysing and evaluating the obtained results both syntactically and semantically and comparing them with the human creations.

Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2015, pp. 125–132, 2015.

2 Related Works

According to our knowledge, not much attention had been paid to this sort of research in Czech context, thus it is difficult to offer a methodological as well as technical comparison of solving similar problems. The text by J. Novotný[1] is a general and philosophical talk which does not contribute much to the considered area.

An attempt to deal with the Czech rhymes was created by P. Šrubař[2], it is interesting, however, it does not touch the grammar and the rules for a metre. There is an interesting text by J. Materna[3] which informs about applying machine learning techniques to this issue. We also would like to mention the text *Poetry, computer and poet* [9] which contains an overview of the experiments in the field.

Other projects concerning generation of poems include the famous Raymond Queneau 100,000,000,000,000 sonnets, or many attempts that generate poems from patterns (e.g. AI poems[4]). Works similar to ours are rather rare but exist, e.g. Automatic Poetry Generator[5].

3 Lexicon: Corpus of Czech Poetry

Compared to the early work [4], we have immense possibilities of computational power and storage. In the early work, the authors created the lexicon from some writings by Jaroslav Seifert. Now we have decided to create the lexicon automatically. We created a 2 million corpus of Czech poetry (and Czech translations of non-Czech poetry) from Wikisources[6]. The corpus has been annotated using tokenizer `unitok`, and the tagger `desamb` [5]. Afterwards, we let the Sketch Engine [3] compute word sketches and exported the most frequent 1000 words.

It can be seen that the language of poems differs significantly from the language in general corpora. Among the 1000 most frequent words from the poetry corpus and the big web corpus `cztenten12`, only 423 word were in both corpora. To compare, the same comparison between `cztenten12` and another Czech corpus `Czes2` founds 817 common words.

"A word sketch is a one-page, automatic, corpus-derived summary of a word's grammatical and collocational behaviour" [1]. In the poetry corpus, the word sketches differ a lot compared to the `cztenten12` corpus. Not surprisingly, the distributional thesaurus for the poetry corpus differs from the

[1] `http://www.fi.muni.cz/usr/jkucera/pv109/2004/51914-pocitace_a_poezie.html`

[2] `http://www.rymy.cz/about.htm`

[3] `http://www.mlguru.cz/stredovek-umele-inteligence-skoncil-seznamte-se-s-neuronovymi-sitemi-ktere-umi-psat-basne/`

[4] `http://www.aipoem.com`

[5] `http://www.languageisavirus.com/automatic_poetry_generator.html`

[6] `https://cs.wikisource.org`

thesaurus in `cztenten12`. We benefit from the thesaurus when building a stanza from the verses (see Section 5).

4 Formal (DC) Grammar Generating Czech Sentences

Similarly to the previous work, we focused on Czech iambic verse. In Czech poetry, trochees dominate, iambic metre is somewhat unusual because the stress is always on the first syllable. Thus, Czech iambic verses have to start with monosyllabic words but in Czech, most words are longer than one syllable. Nevertheless, many Czech poets wrote iambic poetry or combined verses (e.g. alternation of 5-feet and 4-feet iambic verse or alternation of iambic and trochaic verse). In order to generate Czech verses with a given metre we have to deal with Czech grammar, verb valencies, and poetic rules.

4.1 Grammar for Czech Sentences

The language structure, i.e. grammar: in our case, we decided to work with a small definite clause grammar (DCG) [7] of Czech named Klara and developed in the NLP Centre. It is a formal device able to recognize (and generate) Czech sentences. We adapted a subset of Klara grammar in order to generate correct Czech sentences. The subset concerns noun phrases (with adverbial and adjective modifiers), pronominal phrases, verb phrases (incl. compound verbs), and simple adjective phrases.

In Czech we have to deal with two types of obligatory grammatical agreement that are satisfied in the grammar:

- grammatical agreement in gender, number and case among all components of a noun or pronominal phrase, and between the subject
- grammatical agreement in gender and number between the subject and the verb in past tense or conditional

We also included interjections and particles. First, they appear in poetry (probably more often than in other texts), second, many interjections and particles are monosyllabic words and therefore are useful as first words in iambic verses (see Section 4.3). Using the grammar outlined in this Section, we generated over 1,8 millions of verses.

4.2 Verb Valencies

The rules of our DC grammar also allow us to capture partly semantic aspects of the generated sentences using valency frames (VerbaLex [2]). In this way, we may control the meaningfullnes of the generated sentences in a reasonable extent. DCG produces not only Czech sentences but also their syntactic representations in the form of syntactic trees. In the preceding experiment with generating Czech verses, a fragmentary context-free grammar was used [6]. This time, we have decided to work with a DC grammar which is

more adequate for handling natural languages (like Czech). We extracted right verb valencies (corresponding to the object) from all verbs in the lexicon using VerbaLex. Although a verb can have several right valencies, we took only the first for each verb synset. For this reason, our system cannot generate sentences such as *Peter went from Prague to Brno* but only *Peter went from Prague* and *Peter went to Brno*.

4.3 Poetic rules for the Czech iambic verse

The poetic rules working on the language material, i.e. on the Czech sentences produced by the DC grammar mentioned above were implemented directly as a DCG constraint. In the previous work [6], the poetic rules were applied to generated phrases.

First, we have to define some basic notions. A syllable is a group of letters containing one vowel or diphthong or one syllabic consonant, i.e. r or l in the environment `CrC` or `ClC`. A word with $1 \ldots n$ syllables is a word containing $1 \ldots n$ vowels. The Czech vowels are the following: a, á, o, ó, u, ů, ú, i, í, e, é, y, ý, ou, au, eu plus the syllabic consonants r, l (also m, n could be considered but we leave them out here).

Each syllable (vowel) in a verse is considered to be a position (symbollicaly denoted by p, each position is either even or odd. A null position, i. e. a position immediately before the first occupied position, is automatically treated as an even position. For generating an N-syllable iamb (a $N/2$-feet iambic verse), we assume that we have N occupiable positions.

The following rules of word selection and rules regulating the number of syllables can be formulated:

1. If the last occupied position is even, i. e. if $p = 0, 2, 4, \ldots$, it is necessary to select a monosyllabic word. This means that each verse has to start with a monosyllable.
2. If the last occupied position is odd, i. e. $p = 1, 3, 5, \ldots$, it is necessary to select a $2 \ldots n$-syllabic word whose maximum number of syllables follows from the relation $n = N - p$ (without enjambments $n = N - 1 - p$).

2b. A weaker formulation: if the last occupied position is odd, i. e. if $p = 1, 3, 5, \ldots$, it is possible to select any word X, even a monosyllable, whose maximal number of syllables follows from the relation $n = N - p$ ($n = N - 1 - p$). This rule, if used, would cause a considerable loosening of the rather rigid poetic rules.

Rules for rhymes such as `aa`, `bb`, `cc` ... or `ab ab cd cd` ... or `abba`, `cddc` ... in a iambic verse can formulated as well [6] but we plan to apply them in the future research.

In contrast with the previous work, we implemented the poetic rules directly to the DC grammar. In order to generate mostly sentences that correspond to the poetic rules, we start each sentence with a monosyllabic word and afterwards, we generate n-syllable iambs for $n \in 5, 8, 9, 10, 11, 12$. The verses

have to have stress at least on some even syllables, so the program generates words longer than two syllables. This approach is close to the rule 2b.

In the second stage, we combine the generated verses in order to create a stanza. We observed the stanzas in poems by Czech poets Boleslav Jablonský (19th century, early romantism), Karel Sabina (19th century, romantism), and František Hrubín (20th century, modernist) and copied their schemes. For example in Hrubín's poem Milostná, the stanza contains verses of 8, 11, 5, 8, 11, 5 syllables, most of them purely iambic.

5 Adding Semantics to the Generated Verse

The combination of generated verses is partially random but constrained by the poetic rules for Czech iambic verses and also by the patterns of stanzas. These constraints can lead into non-sense combination such as illustrated in Example 5. No semantic relation among the words větvím (branches), mořím (seas), hudbám (music), hadovi (snake) is apparent. For this reason, we employed the Sketch Engine Thesaurus [8] in order to keep a particular topic for most of the verses in the stanza. Such interconnected verses that form a stanza are shown as Example 5.

To conclude, the generated verses are constrained not only by grammatical (agreements), syntactical (word order, verb valencies), and poetical (number of syllables, stress) rules but also partially by semantic rules. For this reason, we need a very large number of verses from which the program can select the appropriate ones.

co bere celým starým větvím kde
jen dám si novým mořím
co bere celým bílým hudbám ven
sem bijeme hadovi

Fig. 1. A stanza without semantic information.

ty výš běžela do kruhů
co pryč bijeme **obrazům**
kde dám **krásе**
co kdes bijeme **otcovi**
co hodně bijeme věci
též dám **krásám**

Fig. 2. A stanza with the topic víra (belief), words similar to the topic are marked in bold.

Sketch Engine Thesaurus for víra (belief) : **obraz** (**image**), pravda (truth), církev (the Church), vůle (will), domov (home), myšlénka (thought), právo

(right), milost (compassion), cit (feeling), rozkoš (delight), místo (place), hřích (sin), svoboda (freedom), štěstí (happiness), vlast (homeland), řád (order), mír (peace), vděk (gratitude), řeč (speech), víska (small village), naděje (hope), radost (joy), . . . **krása** (**beauty**), nevěsta (bride), čest (honor), **otec** (**father**) . . .

6 Conclusions and Future Work

We have built a Prolog DC grammar that generates Czech iambic verses using a thesaurus-based stanza builder. The program can be used via the web interface at `https://nlp.fi.muni.cz/projekty/czech_verse/`. The quality of the resulting poetry is going to be a subject of a future poetry evaluation in a cooperation with the versologists.

So far, we have not implement rhymes in the stanza builder. This feature will be developed in the future.

In the paper, we have been trying to model a situation, in which a poet – human being – selects some syntactic structures and refuses other because of the inconsistence with his poetic intentions. The analogy between the poet-creator and a computer is rather superficial – the computer's creative intentions are dependent on our ability to formulate as adequately as possible the creative properties of a poet. We have to realize that the machine is able to solve the conflicts using combinatoric techniques only – by the systematic search of all existing variants until the first acceptable one is found.

One important remark has to be made: in the paper, we have not paid much attention to the semantic aspects of the poetic creation though they play a relevant role in this respect. The semantics influences strongly a poetic message that poets try to express in their verses. Not speaking about integrating emotions into poetic creations. In this respect, our possibilities are quite limited so far even though an existing research in the area of the affective computing offers some progress. But this is a topic for another paper.

Acknowledgement This work has been partly supported by the Masaryk University within the project *Čeština v jednotě synchronie a diachronie – 2015* (MUNI/A/1165/2014) and by the Ministry of Education of CR within the Czech-Norwegian Research Pogramme in the HaBiT Project 7F14047 and within the LINDAT-Clarin project LM2010013.

References

1. Word sketch | Sketch Engine, `https://www.sketchengine.co.uk/word-sketch/`, [accessed online, 2015-11-01]
2. Hlaváčková, D., Horák, A.: Verbalex – new comprehensive lexicon of verb valencies for Czech. In: Proceedings of the Slovko Conference (2005)
3. Kilgarriff, A., Rychlý, P., Smrž, P., Tugwell, D.: The Sketch Engine. In: Proceedings of the Eleventh EURALEX International Congress. p. 105–116 (2004), `http://www.fit.vutbr.cz/research/view_pub.php?id=7703`

4. Levý, Jiří, Pala, K.: Generování veršů jako problém prozodický [Generating verses as a prosodic problem]. In: Palas, K.; Levý, J. (ed.) Teorie verše. II, Sborník druhé brněnské versologické konference, 1967
5. Šmerk, P.: K morfologické desambiguaci češtiny [Towards morphological disambiguation of Czech]. thesis proposal, Masaryk University (2008), http://is.muni.cz/th/3880/fi_r/
6. Pala, K.: Conflicts between grammar and poetics. In: Prague Studies in Mathematical Linguistics IV. pp. 229–240. Czechoslovak Academy of Sciences (1973)
7. Pereira, F.C.N., Warren, D.H.D.: Definite Clause Grammars for Language Analysis - A Survey of the Formalism and a Comparison with Augmented Transition Network. Artificial Intelligence pp. 231–278 (1980)
8. Rychlý, P., Kilgarriff, A.: An efficient algorithm for building a distributional thesaurus (and other sketch engine developments). In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. pp. 41–44. ACL '07, Association for Computational Linguistics, Stroudsburg, PA, USA (2007), http://dl.acm.org/citation.cfm?id=1557769.1557783
9. Uličný, O.: Poezie, počítač a básník [Poetry, computer and poet]. Čeština doma a ve světě 5(3), 27–32 (1997)

Examples of computer generated poetry (Czech)

ty tady běžela do věr
co bere nebesům místo
tak sem chvěje mojí krásou
co ráno bijeme hříchu

on, jenž byl svět, byl co milejší ret
co bere celým krásným rakvím tíž
on, jenž byl svět, byl ty dobrý věk
co bere celým novým slastem ráz
tak tak čekají jeho malého
on, jenž byl svět, byl on tichý bez

tak též cítila se neplné časy
ty hlouběji běžela do časů
co nic bijeme celým bílým zářím
tak tu cítila se nevětší ženu
tak ještě cítila se plný den
on, jenž dal srdci, dal si drahým stromům
tak nedaleko čekají mé rety
dnes bijeme hradům

co bere celým jiným věžím nic
ty nejlépe běžela do ramen
tak hodně chvěje jeho horou
co bere celým zlatým dveřím nic
ty někdy běžela do matičky

tak radší chvěje její horou

on, jenž byl svět, byl on velký les
co bere celým velkým stráním hoře
tak radši cítila se mladší jiné
co bere celým starým hudbám klidně
co bere celým zlatým růžím sladce
on, jenž dal srdci, dal nám mladým jiným