Intelligent Search and Replace for Czech Phrases

Zuzana Nevěřilová and Vít Suchomel

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
{xpopelk,xsuchom2}@fi.muni.cz

Abstract. This work proposes a new improvement of the 'Search and Replace' function well known from most text processing software.
The standard search and replace function is used to replace exact form of words or phrases by another words or phrases in text documents. It is quite sufficient for languages with minimal inflection such as English. However, a well working word or phrase replacement function for morphologically rich languages requires much more thought.
We explore the issues of implementing a useful search and replace in the Czech language and propose solutions to majority of the problems: A syntactic parser is employed to identify the phrases containing the search word or phrase. The correct word forms used as a replacement are generated by a morphological analyser.
A web demonstration utilizing the proposed solution is presented. The attached examples of use reveal the cases in which the implemented method works well.

Keywords: search and replace, detecting phrases, generating phrases, subject-predicative complement

1 Motivation

This work introduces an intelligent search and replace editing tool for a text processor in Czech. The basic search and replace is a well known function implemented by most text processing software. A search word (or a phrase) and the new text to replace by is entered by the user. The text processor finds all occurrences of the search string and performs the replacement. Exact matches are made therefore the user has to know exact forms of the search phrase. (Some tools support searching regular expressions but that does not offer any grammatical advantage over the basic method.)

The basic function is quite sufficient for languages with minimal inflection such as English. Nevertheless, global search and replace without additional knowledge can cause problems (see e.g. the Scunthorpe problem[1]).

We call our tool "intelligent" since it uses language knowledge to 1) search all occurences of a word regardless its forms, and 2) avoid searching

[1] See https://en.wikipedia.org/wiki/Scunthorpe_problem

Aleš Horák, Pavel Rychlý (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2014, pp. 97–105, 2014.

coincidentally equal substrings. Our work deals with the issue for Czech – a highly inflective language. The improved function should be able to find all inflected forms of the search phrase and apply the respective morphological forms to the replacement phrase. The new search and replace tool offers the ability to automate tedious manual edits like the basic function. A big concern should be placed in reaching a reasonable precision. The users would not tolerate it making mistakes.

Having such tool, one could automate many frequent cases of replacement to save editing time:

- correction of often repeated mistakes,
- unification of terms in translation (done by the chief translating editor),
- especially unification of terms in localisation (e.g. GUI descriptions: 'dialog box' – 'dialogové okno' (neuter gender), 'dialogový box' (masculine inanimate),
- adjusting general parts of manuals to particular products,
- changing ingredients in recipes,
- replacing person or company names in standardised documents,
- especially in legal text, e.g. common parts of contracts
- and other standardized labels, notices, signs.

2 Related Work

There is no work dealing with search and replace function of Czech phrases known to the authors in the Czech speaking environment. Although there is e.g. a Czech grammar checker available for Microsoft Word [6], arguably the most widespread text processor used for Czech, there is no module for search and replace available in the tool.

A general idea of morphological search and replace has been patented by Microsoft [8]. In addition to that, the approach presented in this paper covers also search and replace of multiword phrases and is able to deal with Czech phenomena that are uncommon to English – grammatical gender and subject-predicative complement agreements.

3 Methods

In order to replace whole phrases, we need to identify them in many different forms. For this reason, we use the syntactic parser SET [2]. For lemmatization and tagging, we use the tools `majka` and `desamb` respectively. The parser takes tagged vertical as input, separates individual phrases and determines their heads. It also determines the *phrase lemma* and *phrase tag*: in case of noun phrases, the phrase tag corresponds to the head tag; in case of prepositional phrases, it corresponds to the complement noun phrase. The phrase tag determination is crucial for the replacement method. The phrase lemma corresponds to the respective lemmata in the phrase but in addition, adjective, pronoun, and

numeral modifiers are changed to fulfil the grammatical agreement. For example, the noun phrase "tato dvě červená jablka" (these two red apples) has the respective lemmata: "tento" (this), "dva" (two), "červený" (red), "jablko" (apple). The noun phrase is at the same time the noun phrase lemma since it is nominative.

If the replacement concerns head of a noun phrase or head of a complement noun phrase (in prepositional phrases), the head tag gender and number is compared to that of the new phrase. The replacement has to follow the same case and number as the original phrase tag. The gender can be different and if it is, the obligatory agreements have to be fulfilled.

In Czech, three basic types of grammatical agreement are related to noun phrases:

- head modifiers: this grammatical agreement does not occur solely in Slavonic languages, it is also known in Romance languages: adjectives, pronouns and numerals modifying a particular head have to agree in number and gender with the head
- active verb: the grammatical agreement between the subject and the verb phrase can be complicated for analytical verbs (e.g. past tense in Czech, that is composed from active verb *to be* in present tense and past participle), moreover it relies on a correct detection of the syntactic subject
- predicative complement: if the complement is an adjective, pronoun or numeral, it has an obligatory agreement in gender and number with the subject. In this case, the predicative complement is hard to detect (it depends on the copula verb occurence and the copula verbs are defined in a very arguable way).

More formally, the replacement of phrase p by r in text T ($p \rightarrow r$) proceeds in the following way:

1. detect phrase lemma $p(lemma)$ and phrase tag $p(tag)$ for search phrase p, and phrase lemma $r(lemma)$ and phrase tag $r(tag)$ for the replacement r
2. parse whole T, separate individual noun phrases and prepositional phrases, detect their phrase tags and phrase lemmata
3. if $p(lemma)$ is found in i-th phrase in T:
   (a) replace $p_i(lemma)$ with $r(lemma)$, we label this particular replacement by the same index: r_i
   (b) modify gender of all adjective, pronoun, and numeral modifiers of $r_i(lemma)$ if the genders of $p(tag)$ and $r(tag)$ differ.
   (c) if $p_i(tag)$ is not nominative, decline $r_i(lemma)$ according to the case of $p_i(tag)$
   (d) if r_i is part of the subject and the clause contains a copula verb, modify the predicative complement according to the gender of $r_i(tag)$ (only adjective, pronoun, and numeral predicative complements are subject of gender agreement)
   (e) if r_i is part of the subject and the verb phrase contains a participle, modify the verb participle according to the gender of $r_i(tag)$

For all mentioned types of inflection, we use the declension and conjugation tool [5] based on morphological generator majka [7]. Examples of grammatical agreements follow:

Adjective Modifiers vysoký dům → vysoká budova

Subject-Predicate dům stál na nabřeží → budova stála na nábřeží

Subject-Predicative Complement dům byl vysoký → budova byla vysoká

One imporant pitfall concerning the declension exists: the tagger can detect incorrectly the case or the gender of the noun phrase/prepositional phrase. For example, for the noun phrase "zahraniční podnik" (foreign enterprise) the case can be either nominative or accusative (the word forms are the same). In this case, the parsing passes without problems with both nominative or accusative but the quality of the tagging has serious consequences. If we replace "podnik" (enterprise, in Czech masculine inanimate) by "firma" (company, in Czech feminine), the resulting forms differ depending on the case ("firma" in nominative and "firmu" in accusative). In addition, other sentence parts (verb phrase or predicative complement) may or may not change (depending on whether the phrase is a syntactic subject).

The replacement is therefore "intelligent" in the sense that it replaces noun phrases and prepositional phrases not substrings. Nevertheless, is does not take word senses into consideration. This can be an issue when the assumption *one sense per discourse* is not fulfilled. Particularly, in case of phrases that are part of phrasemes, the replacement can lead to incorrect results. For example, imagine replacing "hand" by "foot". In this case, the phrase "on the other hand" will result into "on the other foot". Replacing all occurences regardless the context can lead into errors, however, [1] proved that in 98% cases, the assumption *one sense per discourse* is correct.

4 Results

A web demonstration utilizing the proposed solution for Czech has been made available at `http://nlp.fi.muni.cz/projects/phrase_replace`. We have used the application to perform replacements in instruction manuals, cooking recipes and a definition page from an encyclopedia. The method works well in these cases since there are not many complicated sentences there and the present tense is usually used. See the appendix for comparison of input and output example texts representing these kinds of replacements.

We also tried general replacements of single words as well as multi word phrases. Despite the tool made mistakes outlined in the previous chapter, it performed fairly well in two thirds of cases: 17 of 30 instances of changing "dům" to "stavba" in 25 random sentences from a Czech web corpus were completely right, 4 replaces introduced a small error but the declension of the

target word was right (e.g. ”v stavbě“ instead of ”ve stavbě“) and 9 mistakes were made in declension or in recognising the right part of speech (e.g. ”domů“ can be a noun or an adverb). Pronominal anaphora resolution proved to be an issue in longer and complex sentences.

The accuracy of the method is not perfect. The problematic coverage of the anaphora related issues or word sense desambiguation makes it even harder. Therefore we recommend to ask the user to confirm each replacement of the search phrase and allow a manual edit at key places in the case of utilising the tool in a word processor.

5 Conclusion and Future Work

The paper presents a preliminary work concerning inflection-aware search and replace of phrases in Czech. It seems that not many previous projects pursue this topic not only in Czech but also in other languages. The reason is that a successful replacement tool depends on other NLP tasks. In the current version, we employ morphological analysis, tagging, parsing, and morphological generation for inflection of the replaced phrases as well as phrases that are subject of obligatory agreements. However, the tool is not aware of co-references in the text. Future work will therefore concern three main subtasks of co-reference resolution:

- **Zero subject resolution** In Slavonic languages, the subject does not have to be expressed in each sentence. Zero subject resolution is needed in sentences containing past participles in verb phrases or predicative complements. It is solved e.g. by [4] but not yet used in our tool. Zero subject resolution is useful for replacements such as Bob $\rightarrow$ Alice as subject. For example: Bob se narodil v Brně, kde také vystudoval. (Bob is born in Brno where he also studied.) Alice se narodila v Brně, kde také *vystudovala*. (Alice is born in Brno where she also *studied*.)
- **Pronominal anaphora resolution** Resolution of possessive, personal, and demonstrative pronouns seems to be a simpler task in morphologically rich languages with several grammatical agreements than in English. It is partially solved by [4,3], however not yet implemented in our tool. Pronominal anaphora resolution is suitable for replacements such as batoh $\rightarrow$ taška (backpack $\rightarrow$ bag). For example: Rozdělal jsem přezky na batohu a začal vybalovat věci. Najednou z *něj* vypadl plátěný pytlík s něčím omamně voňavým. (I opened the backpack and started to unpack it. Suddenly, a small canvas sack with something perfumed dropped out of *it*.)
- **Abbreviated forms** Even in technical text where synonymy is not desired, abbreviated forms exist. To our knowledge, no language tool for recognizing abbreviated forms of noun phrases exist. In future, we need to build such tool, probably based on similarity search and/or corpus-based thesauri. Abbreviated forms could correctly replace phrases such as: Elektrické travní sekačky jsou ideální na udržování menších travnatých ploch.

Vyžadujete-li tichý chod a ohleduplnost k životnímu prostředí, vyberte si *elektrickou sekačku*! (Electric lawn mowers are suitable for maintenance of smaller areas. If you require silent operation and environment considerations, choose an *electric mower*!)

Acknowledgements This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013 and by the Czech-Norwegian Research Programme within the HaBiT Project 7F14047.

References

1. William A. Gale, Kenneth W. Church, and David Yarowsky. One sense per discourse. In *Proceedings of the Workshop on Speech and Natural Language*, HLT '91, pages 233–237, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics.
2. V. Kovář, A. Horák, and M. Jakubíček. Syntactic analysis using finite patterns: A new parsing system for Czech. In *Human Language Technology. Challenges for Computer Science and Linguistics*, volume November 6–8, 2009, pages 161–171, Poznań, Poland, 2011.
3. Lucie Kučová and Zdeněk Žabokrtský. Anaphora in czech: Large data and experiments with automatic anaphora resolution. In Václav Matoušek, Pavel Mautner, and Tomáš Pavelka, editors, *Proceedings of 8th International Conference on Text, Speech and Dialogue, TSD 2005*, volume 3658 of *Lecture Notes in Computer Science*, pages 93–98. Springer Berlin Heidelberg, 2005.
4. Václav Němčík. Saara: Anaphora resolution on free text in Czech. In Aleš Horák and Pavel Rychlý, editors, *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2012*, pages 3–8, Brno, Czech Republic, 2012. Tribun EU.
5. Zuzana Nevěřilová. Declension of Czech noun phrases. In Jan Radimský, editor, *Actes du 31e Colloque International sur le Lexique et la Grammaire*, pages 134–138, České Budějovice, 2012. Université de Bohême du Sud à České Budějovice (République tchèque).
6. K. Oliva, V. Petkevič, and Microsoft s.r.o. Czech grammar checker, 2005.
7. Pavel Šmerk. Unsupervised learning of rules for morphological disambiguation. In Petr Sojka, Ivan Kopeček, and Karel Pala, editors, *TSD*, volume 3206 of *Lecture Notes in Computer Science*, pages 211–216. Springer, 2004.
8. J. E. Walsh and R. A. Fein. Morphological search and replace, 02 1999.

Appendix: Example inputs (left column) and outputs (right column) of the intelligent search and replace tool

Example 1: Instruction manual of a tool. Replace *kráječ* → *bazuka*.

Kráječ na potraviny je zařízení, které doma můžete používat na krájení potravin na tenké plátky jako z lahůdkářství. Tyto domácí kráječe fungují stejným způsobem jako komerční, avšak nejsou tak výkonné. Kráječ na potraviny Vám umožní nakrájet maso a sýry na požadovanou tloušťku. Protože každý kráječ je trochu jiný, podívejte se do manuálu na specifickou konstrukci Vašeho kráječe. Většinu kráječů koupíte z velké části složenou. Postavte základnu kráječe na neklouzavý povrch. Vyberte si místo, kde budete mít dost prostoru na práci s kráječem, avšak kde nikomu nebude překážet a kde se nikdo nezraní o jeho velmi ostrý kotouč.

Bazuka na potraviny je zařízení, které doma můžete používat na krájení potravin na tenké plátky jako z lahůdkářství. Tyto domácí *bazuky* fungují stejným způsobem jako komerční, avšak nejsou tak *výkonné*. *Bazuka* na potraviny Vám umožní nakrájet maso a sýry na požadovanou tloušťku. Protože *každá bazuka* je trochu *jiná*, podívejte se do manuálu na specifickou konstrukci *vaší bazuky*. Většinu *bazuk* koupíte z velké části *složenou*. Postavte základnu *bazuky* na neklouzavý povrch. Vyberte si místo, kde budete mít dost prostoru na práci s *bazukou*, avšak kde nikomu nebude překážet a kde se nikdo nezraní o *jeho* velmi ostrý kotouč.

Example 2: A cooking recipe. Replace *prsíčka* → *kýta*.

Troubu předehřejte na 190 °C. Kůži na prsíčkách dobře propíchejte vidličkou, přelijte vroucí vodou a pak nechte dobře oschnout. V těžké, silnostěnné nepřilnavé pánvi pomalu opékejte prsíčka nejdříve kůží dolů. Pak otočte a nechte opéci z druhé strany. Vyjměte z pánve a osolte. Jablka rozložte do lehce olejem vytřené zapékací formy nebo pekáčku. Posypte tymiánem, přidejte svitek skořice a navrch rozložte opečená prsíčka kůží nahoru.

Troubu předehřejte na 190 °C. Kůži na *kýtách* dobře propíchejte vidličkou, přelijte vroucí vodou a pak nechte dobře oschnout. V těžké, silnostěnné nepřilnavé pánvi pomalu opékejte *kýty* nejdříve kůží dolů. Pak otočte a nechte opéci z druhé strany. Vyjměte z pánve a osolte. Jablka rozložte do lehce olejem vytřené zapékací formy nebo pekáčku. Posypte tymiánem, přidejte svitek skořice a navrch rozložte *opečené kýty* kůží nahoru.

Example 3: A definition page from an encyclopedia. Replace *ptakoještěr* → *slepice*.

Pterodactylus („křídelní prst“) byl rod pterodaktyloidního ptakoještěra žijícího na území dnešního Německa a zřejmě i jinde v Evropě a Africe v období svrchní jury (asi před 151–148 miliony let). Tento létající plaz je jedním z prvních objevených a vědecky popsaných ptakoještěrů vůbec. První zkameněliny byly identifikovány již roku 1784 Cosimou A. Collinim. Řádně vědecky popsán pak byl počátkem 19. století. Byl to dravec, který se živil zejména rybami a jinými malými obratlovci, případně i bezobratlými. K lovu mu sloužily také drobné zuby na okrajích čelistí. Rozpětí křídel pterodaktylů dosahovalo jen kolem 1,5 metru u dospělých exemplářů, patřil tedy mezi menší ptakoještěry.[2]

Pterodactylus („křídelní prst“) byl rod *pterodaktyloidního slepice* žijícího na území dnešního Německa a zřejmě i jinde v Evropě a Africe v období svrchní jury (asi před 151–148 miliony let). Tento létající plaz je jedním z prvních objevených a vědecky popsaných *slepic* vůbec. První zkameněliny byly identifikovány již roku 1784 Cosimou A. Collinim. Řádně vědecky *popsán* pak *byl* počátkem 19. století. *Byl* to dravec, který se živil zejména rybami a jinými malými obratlovci, případně i bezobratlými. K lovu mu sloužily také drobné zuby na okrajích čelistí. Rozpětí křídel pterodaktylů dosahovalo jen kolem 1,5 metru u dospělých exemplářů, patřil tedy mezi menší *slepice*.

Example 4: Random sentences from the Czech web. Replace *dům* → *stavba*.

Věděla, že se v hořícím domě nachází člověk neschopný se vlastními silami dostat ven. Každý dům má padacími dveřmi chráněný vchod obrácený na jih a malá okénka tam, kde zeď domu převyšuje střechu domu sousedního. Každý zákazník má již od začátku stavby jasný manuál co vše má dům obsahovat a může si vše lehce kontrolovat.

Věděla, že se v *hořící stavbě* nachází člověk neschopný se vlastními silami dostat ven. *Každá stavba* má padacími dveřmi chráněný vchod obrácený na jih a malá okénka tam, kde zeď *stavby* převyšuje střechu *stavby sousedního*. Každý zákazník má již od začátku stavby jasný manuál co vše má *stavbu* obsahovat a může si vše lehce kontrolovat.

[2] From Czech Wikipedia: `http://cs.wikipedia.org/wiki/Pterodactylus`.

Example 5: A name phrase. Replace *chráněná krajinná oblast → národní park*.

Budeme se snažit vyhlásit chráněnou krajinnou oblast Prameny Ploučnice. Je nejrozsáhlejší chráněnou krajinnou oblastí v České republice, nabízející množství přírodních rezervací. Jako nejrozsáhlejší chráněná krajinná oblast v České republice nabízela množství přírodních rezervací.

Budeme se snažit vyhlásit *národní park* Prameny Ploučnice. Je *nejrozsáhlejším národním parkem* v České republice, *nabízející* množství přírodních rezervací. Jako *nejrozsáhlejší národní park* v České republice *nabízel* množství přírodních rezervací.

Example 6: A change from a male name to a female name. Replace *Václav Havel → Eva Nováková*.

Václav Havel působil v 60. letech 20. století v Divadle Na zábradlí, kde jej také proslavily hry Zahradní slavnost (1963) a Vyrozumění (1965). 9. července 1964 se Václav Havel po osmileté známosti oženil s Olgou Šplíchalovou. Po vypuknutí Sametové revoluce v listopadu 1989 se Václav Havel stal jedním ze spoluzakladatelů protikomunistického hnutí Občanské fórum a jako jeho kandidát byl 29. prosince 1989 zvolen prezidentem Československa.[3]

Eva Nováková působila v 60. letech 20. století v Divadle Na zábradlí, kde *jej* také proslavily hry Zahradní slavnost (1963) a Vyrozumění (1965). 9. července 1964 se *Eva Nováková* po osmileté známosti *oženila* s Olgou Šplíchalovou. Po vypuknutí Sametové revoluce v listopadu 1989 se *Eva Nováková stala jedna* ze spoluzakladatelů protikomunistického hnutí Občanské fórum a jako jeho *kandidát byl* 29. prosince 1989 *zvolen prezidentem* Československa.

[3] From Czech Wikipedia: `http://cs.wikipedia.org/wiki/V%C3%A1clav_Havel`.