Converting the CQL to the Natural Language

Daniela Ryšavá, Nikol Volková
Faculty of Arts, Masaryk University

Outline

- motivation
- input and output
- query structure
- application description
- conversion process
- future work

Motivation

- started as a project at student workshop
- previous project *Queries in CQL (for SkE)* — for the Czech language
- use: hint in concdesc function in Sketch Engine

Input

- CQL query

[lemma="jazyk"]

Output

- converted query in natural language

Hledaný výraz je slovo, které má základní tvar "jazyk".

translation: *Searched expression is a word with the basic word form "jazyk*".

Query structure

<stru/>containing **[attribute op "value"]** *quant within<stru/>*

- attribute: lemma, word, tag
- op: =, !=
- value: string, regular expression
- *quant*: *, +, ?, {x}, {x,y}, {x,}
- *stru*: *s*, *doc*, *p*

Application description

- Python, version 2
- UTF-8 encoding
- regular expressions, conditions, cycles, nested functions
- adding converted parts of the query to the list
- attributive Czech tagset (Brno)

Conversion process – attribute, op, value

operator: =, !=

- negation at relevant positions in output:

    variable *ne* → "", "ne"

    or "cokoli jiného než" (translation: "*anything other than*")

Conversion process – attribute, op, value

attribute: lemma or word

- value: string of charaters of the Czech alphabet
- value: regular expression
  - detecting .* in the string (converted to Czech output)
  - otherwise the output is a regular expression

Conversion process – attribute, op, value

[lemma="pes"][word!="s.*n"][word="[kl]os"]

Hledaný výraz je slovo, které má základní tvar "pes", následuje slovo nezačínající na "s" a nekončící na "n", následuje slovo odpovídající regulárnímu výrazu "[kl]os".

translation: *Searched expression is a word with the basic word form "pes", followed by a word not beginning with "s" and not ending with "n", followed by a word matching the regular expression "[kl]os".*

Conversion process – attribute, op, value

attribute: tag

- value: string of characters of the Czech alphabet and numbers
- converted according to customizable list of morphological tags

Conversion process – attribute, op, value

[tag!="k2.*"][tag="k1.*c5"]

Hledaný výraz je slovo definované značkou jako cokoli jiného než adjektivum, následuje slovo definované značkou jako substantivum, vokativ.

translation: *Searched expression is a word defined by the tag as anything other than adjective, followed by a word defined by the tag as noun, vocative.*

Conversion process – quantifiers

[tag="k2.*"]+ [lemma!="studovat"]? [tag="k7.*"]* [tag="k1.*"]{1,2}

Hledaný výraz je slovo definované značkou jako adjektivum vyskytující se minimálně jednou, následuje slovo, které má základní tvar jakýkoliv kromě "studovat" vyskytující se maximálně jednou, následuje slovo definované značkou jako prepozice opakující se libovolněkrát, následuje slovo definované značkou jako substantivum opakující se 1–2krát.

Conversion process – quantifiers

[tag="k2.*"]+ [lemma!="studovat"]? [tag="k7.*"]* [tag="k1.*"]{1,2}

translation: *Searched expression is a word defined by the tag as adjective, used at least once, followed by a word with the basic word form other than "studovat", that is optional, followed by a word defined by the tag as preposition, used any number of times,followed by a word defined by the tag as noun, used 1-2 times.*

Conversion process – containing, within

structure: s (sentence), doc (document), p (paragraph)

<s/>containing [lemma!="chytat"][word="lelky"]

Hledaný výraz je věta, v níž se nachází slovo, které má základní tvar jakýkoliv kromě "chytat", následuje slovo "lelky".

translation: *Searched expression is a sentence containing a word with the basic word form other than "chytat", followed by a word "lelky".*

Conversion process – containing, within

structure: s (sentence), doc (document), p (paragraph)

[lemma="lingvista"][word="okno"] within<doc/>

Hledaný výraz je slovo, které má základní tvar "lingvista", následuje slovo "okno", to celé v rámci jednoho dokumentu.

translation: *Searched expression is a word with the basic word form "lingvista", followed by a word "okno", all of this within one document.*

Future work

- nested conditions *[lemma="bez" & tag="k1.*"]*
- regular expressions inside a value *[word="[kl]os"]*
- correct tag notation (not *k1c1*, *c1k1*)
- other advanced operators (e.g. *meet*, *union*)

supported by the Masaryk University within the project Čeština v jednotě synchronie a diachronie – 2015 (MUNI/A/1165/2014)

Thank you for your attention.

https://nlp.fi.muni.cz/projekty/cql2cz/