Example use cases for federated content search for FCS workshop in Copenhagen

Maciej Piasecki, Marek Maziarz, G4.19 Research Group, Wrocław University of Technology

A minimal use case: a simple scenario that is still useful to the end users;

Extracting statistics from corpora: a corpus browser should return precise statistics of the search results. The user is interested in comparing occurrences of a given proper name in texts of different languages. For instance, sociologists may be interested in comparing frequences of a given country name/given person name (measure of importance for a community) in different corpora.

Filtering search by metadata (author, date, style, theme, title, genre, source publication):

- scientists interested in language evolution are interested in corpora containing dates of text publication – they could localize linguistic facts in time.
- specialists looking for a text written by a particular author may want to chose texts of the author in one language or in several languages

Finding the *k* most frequent words in a corpus in a defined period of time. The presentation of the results should include comparison of frequency histograms for different words.

Extracting from a speech corpora fragments corresponding to words specified in the query and found in the text transcription of the speech recordings. Presentation of the found fragments in terms of audio concordances: audio recording of the found word together with several preceding and following words.

Finding occurrences of a specified sequence of phonemes (diafons, trifons) in a speech corpus. The presentation of the search results should be encompass: speech recordings, orthographic transcription and phonetic transcription. The results can be further filtered with meta-data concerning the type and source of recordings.

Extracting concordances for a specified sense of a word. The sense will be specified by reference to a specific lexico-semantic resource e.g. a wordnet.

Finding alignment on the level of sentences or even words for a search word (or a phrase) in a parallel bilingual or multilingual corpora, possibly restricted to documents of a specific style, genre, domain etc. It should be possible to distinguish between alignments produced automatically and made manually.

Finding and comparing occurrences of lexical translations for a word across corpora of different languages: Some users may be interested in finding out properties of corresponding verb semantic frames in two languages. He will enter English *eat* and Polish *jeść* and describe their POS (verb) and analyse thematic roles of the two predicates Presentation of the search results should facilitate comparison of uses in different languages.

Finding words used in or described by specified semantic roles or used with a specified valency frame.

A more elaborate use case: something with higher requirements, e.g.

On the input a user specifies a list of words in different languages and the task is to find all sets of sentences in the respective languages such that the sentences are mutually aligned and include the query words (distributed according to the languages).

Finding words, words sequences and word sets described by regular expression in a corpus which is not morpho-syntactically disambiguated or includes both: complete morphological analyses for words and disambiguation tags (added manually or generated automatically). In a query the user can refer to all morphological analyses of a word, selected analyses (by a query language expression), the disambiguation decision. It should be possible to distinguish between unambiguous (with exactly one interpretation) and ambiguous words, to exclude some interpretations by the negation operator and specify only selected elements of a structured tag.

For a inflectional, free word-order language, finding all adjectives modifying a specified noun. Such an adjective can occur in any position in relation to the noun, and can be separated from the noun by several other words. Identification of the modification can be performed by application of a predefined constraint written in a language of morpho-syntactic constraints or done during pre-processing by shallow parsing.

Finding all occurrences of a lexico-syntactic patterns written in a language of morpho-syntactic constraints. A pattern can specify word (or phrase) types, lexical elements, agreements between the values of grammatical categories etc., e.g. *NP … such as NP*. Patterns can be very flexible and cannot be defined in advance – a user must be free in formulating them.

Finding word occurrences in syntactic subtrees of a specified type or specified by a partial pattern.

Searching for all variants of a proper name in the query.

For an inflectional language, finding all occurrences of a multi-word expression in different morphological form, including discontinuous occurrences.

Finding a sequence of grammatical classes (more specific than Parts of Speech) in corpora of different languages, in such a way that the query is specified once for all languages without the necessity to translate it manually to all different tagsets and query languages.

Finding concordances for all possible translations of a given word. The main difficulty is in an appropriate presentation of the search results.

Searching different semantic lexicons (e.g. wordnets) for a words, its senses and semantic relations that describe it. For a given word, finding a path across links of the specified relations starting in the given word, e.g. a hyponymic path to the top node or a hypernymic paths to the hypernymy hierarchy leaves. Finding changes in the description of a word in the previous versions of the resource.

Finding aligned sentences for words from different languages occurring in texts from the same period.