Example queries for Federated search.

Jan Odijk, Utrecht University, 12 april 2013.

Federated search focuses on search in data. But simple queries on metadata such as

(1) Give me a list of all LRs for the Dutch language
(2) What is the size of all Dutch text corpora (in #tokens)
(3) Give me a list of all Dutch data that contain children 2-7 years old as speaker
(4) Give me a list of all Dutch data containing any of the words *heel*, *zeer, erg*

Are also important. Most of these (pretty simple) queries can currently not be posed at all from a single place, even if we restrict attention to the data hosted at one centre (e.g. INL, or MI). I am not sure that metadata search in CLARIN currently already enables such queries. But CLARIN should make such queries possible and easy

Concerning Federated Search In data:

Assume that for the terms in boldface corresponding DCs in ISOCAT have been defined. And that these DCs are used instead of the boldface terms.

In addition, assume that the resources in which the search takes place has mapped its data categories to ISOCAT DCs.

First, a relative simple query:

Search in WFT-GTB

Give me **entries** with **PoS=noun** of which the **headword** ends in "tsje"

Slightly more complicated

Search in GTB, CELEX, CGN-lexicon

Give me **entries** with **PoS=noun** and with the **headword** ending in "tsje", together with the **source** (=GTB, CELEX, of CGN-lexicon) in which it was found.

Slightly more complicated

Search in all resources where the **language=nld**

For each **resource** with **language=nld**
  For each word in ['zeer', 'heel', 'erg'] with **PoS=adj**
    For each **sense** of the word
      For each **synonym** of the **sense**
        For each **lemma** of the **synonym**
          Return word, **Pos**, **sense**, **synonym, lemma,**
          'synonym' , **resource**.**name**

And analogously with '**synonym'** replaced by **'immediate hyperonym'**
And analogously with **'synomym'** replaced by **'hyponym'** (incl hyponyms of hyponyms)

Question: Will federated search somehow smartly `know' (e.g. from the metadata) that it has to search in lexicons only, actually only in lexicons that contain synonym information? Or will it waist time and effort by searching in all text corpora and in lexicons that do not have synonym information? Or is a smart choice of resources to search in left to the user?

Similarly:
Search in **CGN**
Give me all **utterances** that contain the word 'zeer' with **PoS=ADJ**
Spoken by a **speaker** with **age<=7.**

(there are no speakers with age<=7 in CGN; will federated search smartly be able to see this from the metadata or will it waist time searching?)

More Complicated

Search in the virtual collection consisting of the CGN-corpus, VU-DNC, SONAR, CGN-lexicon, CELEX-lexicon.

Give me **utterances** that contain a subsequence of the form:

- A **wordtoken** with **PoS**='**definite determiner**', immediately followed by
- A **wordtoken** with **PoS**=**adjective** with **vorm**=**zonder-e,** immediately followed by
- A **wordtoken** with **Pos**=**noun**

(examples are '*het bijvoeglijk naamwoord*', '*de gulden snede*', '*het ingewikkelder probleem*')
Alternative: just return the subsequence

Still more complicated:

The same as in the preceding example but now

- the adjective should not end in two syllables that both contain a schwa (represented by a regular expression over the phonetic transcription) in its **phonetic_transcription** as found in the CGN-lexicon (

This excludes an example such as: 'het ingewikkelder probleem'.

Even more complicated:

As in the preceding example, but

- a value for an additional attribute with as possible values **eFormExists, eFormDoesNotExist, eFormExistenceUnknown.** The value specifies whether **it is true for the adjective** that a form with property **vorm**=**met-e** exists or whether it is unknown whether such a form exists.

The latter is determined as follows: ** check this**

- let *wv* be the value of the attribute **word** of the word**token** with properties **Pos**=**adjective**, **vorm**=**zonder-e**). Look up the **entry**/**ies** for *wv* for which PoS=adjective in the CGN-lexicon and/or CELEX-lexicon lexicon, and determine its **lemma** (=*wl*)
    - if not found: result =**eFormExistenceUnknown**
    - if found
        - look up in CGN/Celex an entry with **PoS=adjective**-code and **lemma=wl** and **vorm**=**met-e**
            - if found: result=**EFormExists** (e.g. (het) bijvoeglijk (naamwoord))
            - if not found: result= **eFormDoesNotExist** (e.g. ('de) gulden (snede)'

This can be done in one very complicated query, or the queries might be put in a series where the results of the First query are filtered by the second query, etc.

Iterative Case
Each result in of the previous query is (or contains) a sequence Det ADJ NOUN
For each result found in the previous query,

Give me **utterances** that contain a subsequence of the form:

- A **wordtoken** with **PoS**='**definite determiner**', immediately followed by
- A **wordtoken** with **PoS**=**adjective, with lemma=**ADJ**.lemma and** with **vorm**=**met-e,** immediately followed by
- A **wordtoken** with **Pos**=**noun** with **number=**NOUN**.number**

(alternative: just return the subsequences)

Will Federated content search enable queries of the types and complexity illustrated above?

Other important item: Federated search presupposes semantic interoperability**.** ISOCAT etc enable semantic interoperability, but **only if the structure of similar resources is identical**. If the structure deviates, additional measures are needed. In reality, even closely related resources differ (slightly) in structure (e.g. CGN v. SONAR). This requires attention. I have an (unfinished) document on this topic

For more examples of desired search see

Odijk, J. (2011), "User Scenario Search", internal CLARIN-NL document, April 13, 2011. [docx]