# Topic: Search engine upgrades: extension to treebanks

**IDs in Google Sheets:** T83

**Author:** INT/Jesse de Does

**What exists:** Token level annotation search for large corpora; separate treebank query engines

**What must be adapted / extended / created anew:** Because of richer and more intuitive searching for syntactic phenomena, it is surprising that treebanks have not superseded token-based corpora as the basic facility for corpus retrieval.

In this task, we develop a scalable integrated treebank search and exploitation infrastructure for large corpora, integrated with the BlackLab search engine.

Requirements

* Performance. Performance at least for simple queries comparable to CQL (e.g. /node[@cat="np" and node[@rel="mod" and @pt="adj"] and node[@rel="hd" and @pt="n" and @lemma="schaap"]]) should not be much slower than [pos=”adj”][lemma=”schaap”]
* Scalable, distributable, federable
* Support both dependency and phrase structure
* Enable combination of CQL constructions and treebank query (?)

Benefits:

* Leverage efficiency and flexibility of the lucene Span query mechanism and lucene/solr solutions for scalability and distributed processing to achieve goals
* One infrastructure to publish corpora instead of different copies of e.g. opensonar in whitelab and gretel3, so management of metadata, distributed deployment, user annotation, statistics and visualization, etc, etc need not be duplicated

**Why important for CLARIAH (scientific impact):**

Current corpus exploitation infrastructure for treebanks and token-based corpora is fragmented and not always scalable; it is difficult to integrate analytics from treebanks and token-based corpora.

On the one hand, analytics, query management, analytics and visualization in treebank applications lag behind the facilities offered by mature token-based corpus exploitation applications like the Sketch Engine, IMS, BlackLab Server, WhiteLab and the Nederlab portal. On the other hand, for many tasks (even very simple ones like finding typical subject or object nouns for a given verb), token-based queries are no more than a poor man’s substitute for phrase structure search. A single, scalable mode of access to modern Dutch data will greatly benefit both linguistic and applied research.

**Targeted/Actual users:** linguists; computational linguist; language learners

**Actual use (quantify!):** Since the envisaged facility will have all capacities of token-based and treebank query and more, a satisfactory implementation will attract the attention from both user groups

**Social Impact** **(concrete examples):** Facilitate linguistic research based on corpus data by creating a single scalable retrieval hub

**Estimate in PMs (try to justify):** 8PM. Production-quality implementation of new technology is required; user interface needs to be enhanced; 3PM functional design and testing

**Lead + PMs:** INT: 8

**Participants + PMs:** INT

**Deliverables and milestones:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Task | **ID** | **Class** | **Type** | **Description** |
| T083 Treebanks | T083M1 | Milestone | Document | Document defining the requirements for the scalabale treebank search, based on input from a set of use cases. |
|  | T083D1 | Deliverable | Software | Scalable treebank search implementation base on BlackLab and GreTeL |
|  | T083D2 | Deliverable | Document | Document describing T083D2 and its application to use cases. |