Corpus research with “CorpusStudio”

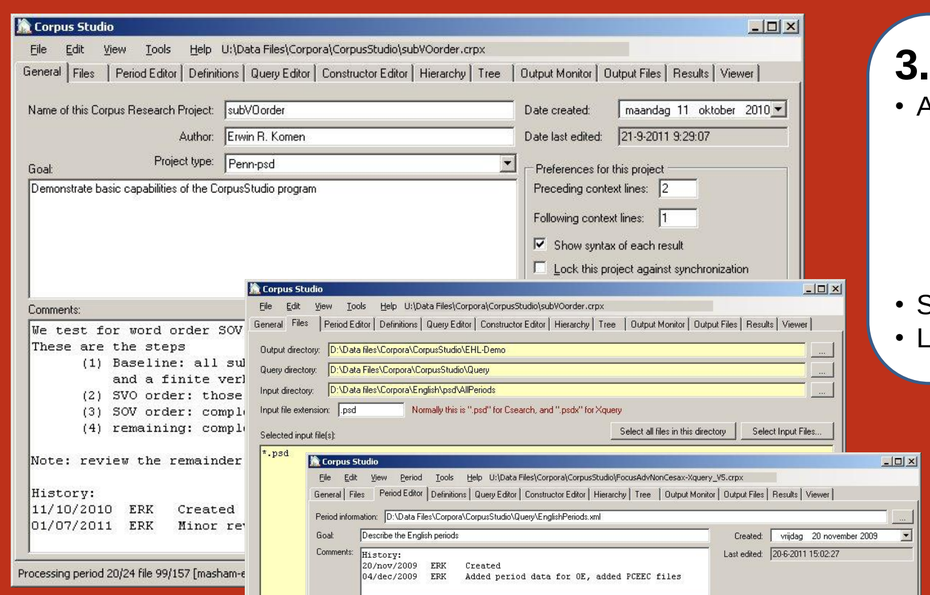
Erwin R. Komen, Centre for Language Studies, Radboud University Nijmegen / SIL-International
E.Komen@Let.ru.nl

1. Background

- Corpus research wish list:
  - Repeatable results
  - Hierarchical queries
  - Windows user interface
- CorpusSearch2 (Randall et al., 2005):
  - Need command-line shell
  - Either type in or know to work with batch files
- TigerSearch, TigerGraphViewer (Corex):
  - Uses windows interface
  - Only works with *xml* stand-off format

2. CorpusStudio goal

- Graphical interface between user and search engine
- Repeatable results
- Hierarchical queries
- Facilitate different formats
  - Treebank (using CorpusSearch2 engine)
  - Cesax output (psdx, using Xquery)
  - Negra, CGN (tiger, using Xquery)
  - Alpino (using Xquery)
- Provide main statistics of results
- Access to individual results + definable context (for information structure research)

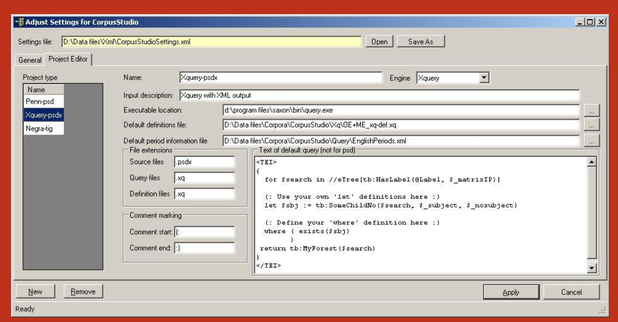


3. Corpus Research Projects

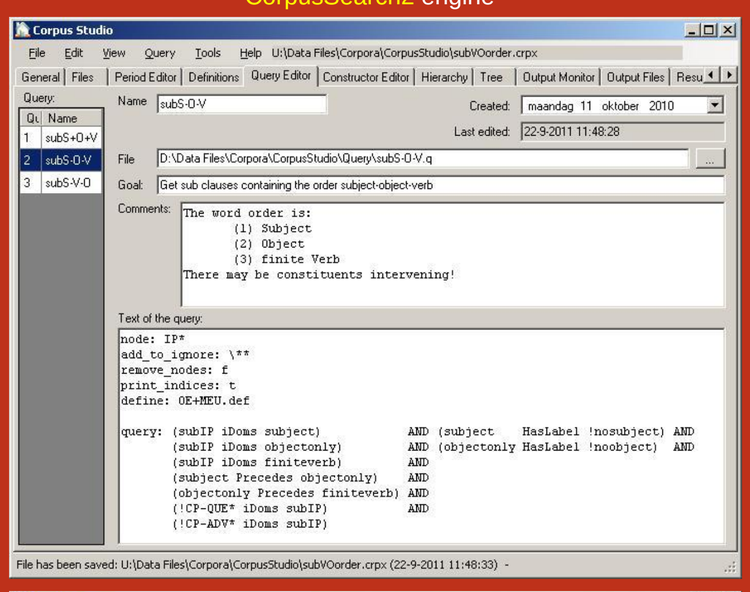
- All information for one research project
  - Meta information (author, dates, goal)
  - Input and output file directories
  - Which input file belongs to which period
  - All definition and query files used
  - Execution order
- Synchronization (periods, queries, definitions)
- Lockable (exchange & repeatability)

4. Defining queries

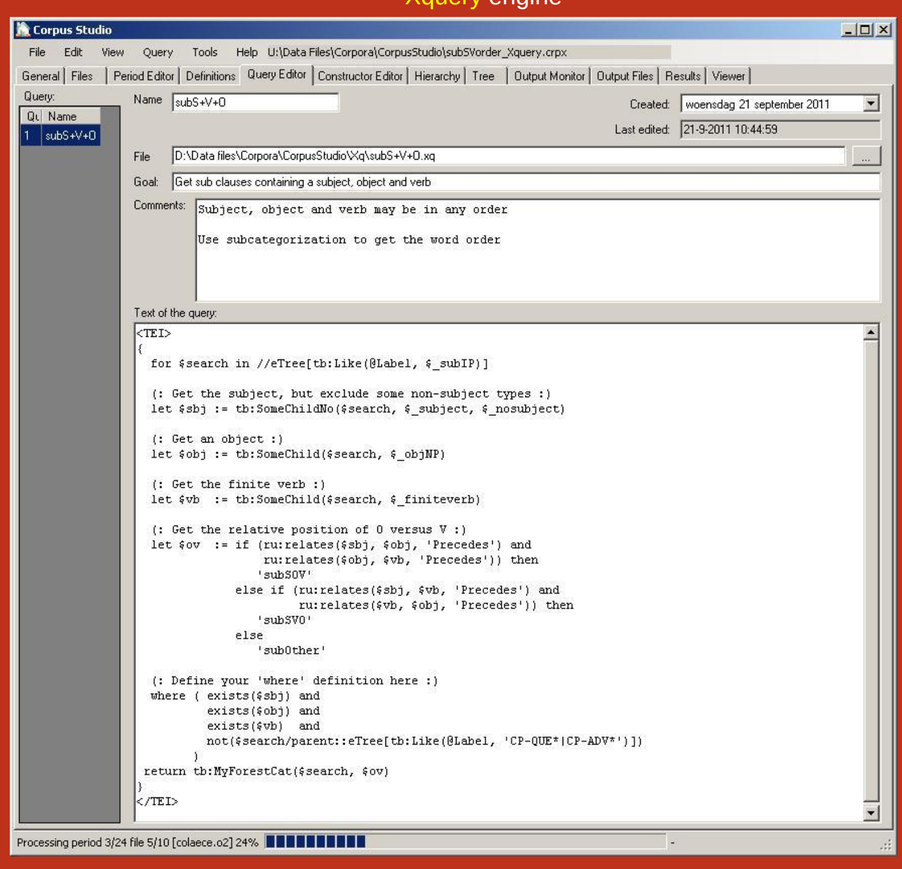
- Definition editor
  - Constants
  - Functions (Xquery)
- Query editor
  - Subcategorization (Xquery)
- Constructor editor
  - Execution order
  - Options (examples, output, complement)
- Verification of query hierarchy



Penn-psd project
CorpusSearch2 engine

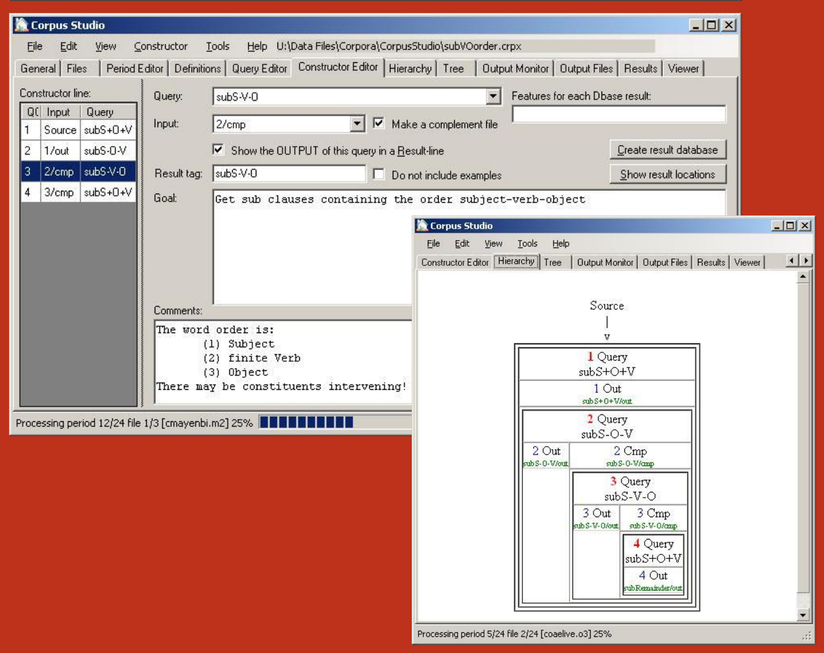


Xquery-psdx project
Xquery engine



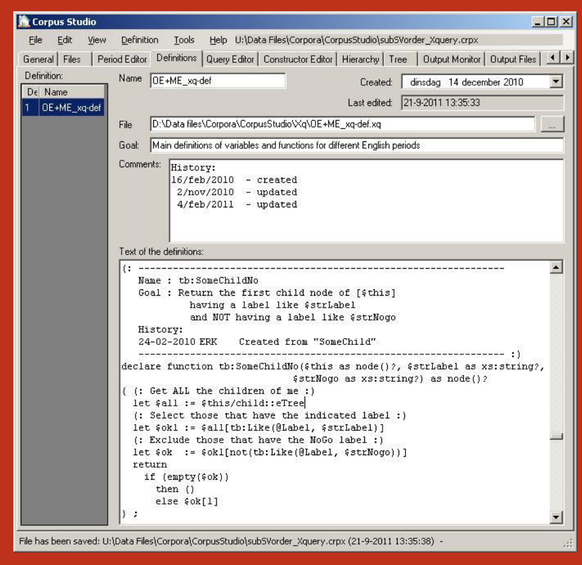
5. Query execution

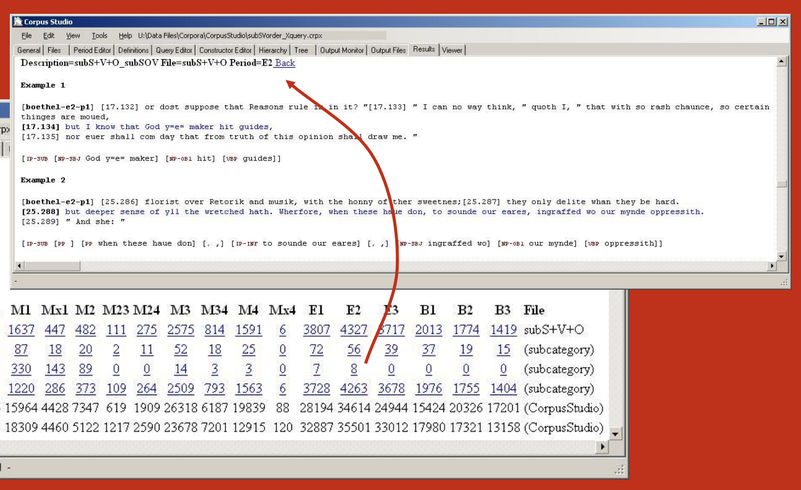
- Early query error detection
- Progress indicator
- Execution order
  - For each $period …
  - For each $text in $period …
  - For each $forest in $text … (Xquery)
  - For each $queryLine in $forest



6. Research project results

- Main statistics overview
- Individual results
  - Definable context (IS research!)
  - With syntax (if requested)
  - With additional information (Xquery)





7. Conclusions

- CorpusStudio facilitates
  - User-friendly interface
  - Repeatable results
  - Students’ participation in corpus research
- The bottom line: it’s for free!
  - http://erwinkomen.ruhosting.nl/software

8. References

Boag, Scott, Don Chamberlin, Mary F. Fernández, Daniela Florescu, Jonathan Robie, and Jérôme Siméon. 2010. *XQuery 1.0: An XML Query Language (Second Edition): W3C Recommendation, <http://www.w3.org/XML/Query>.*

Brants, Sabrin, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. *“The Tiger treebank”,* Proceedings of the Workshop on Treebanks and Linguistic Theories Sozopol: <http://www.ims.uni-stuttgart.de/projekte/TIGER/paper/treeling2002.pdf>.

Randall, Beth, Ann Taylor, and Anthony Kroch. 2005. *CorpusSearch 2, <http://corpussearch.sourceforge.net>.*

Radboud University Nijmegen