

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

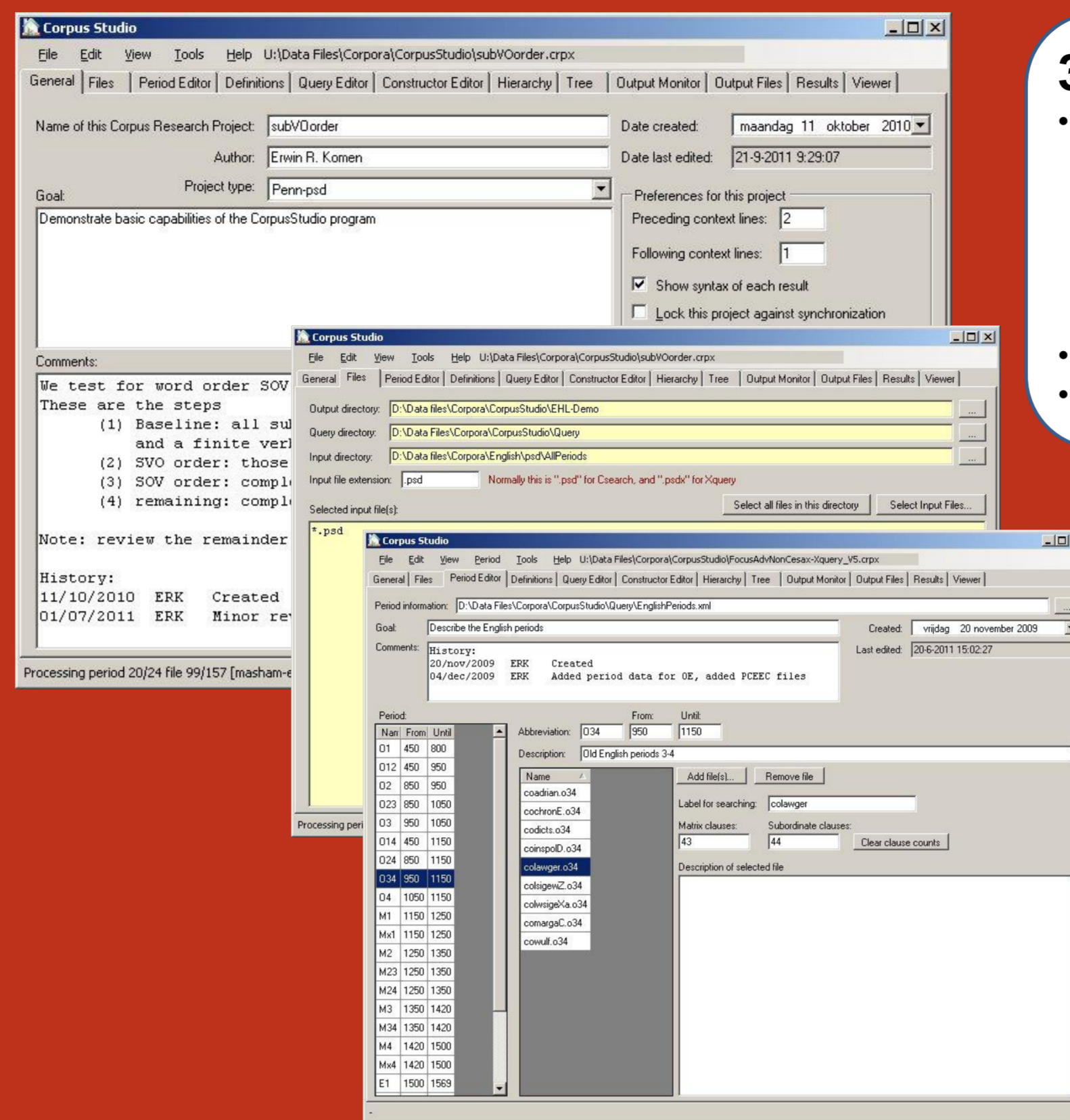
- Graphical interface between user and search engine
- Repeatable results
- Hierarchical queries
- Facilitate different formats
 - Treebank (using CorpusSearch2 engine)
 - Cesax output (psdx, using Xquery)
 - Negra, CGN (tiger, using Xquery)
 - Alpino (using Xquery)
- Provide main **statistics** of results
- 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

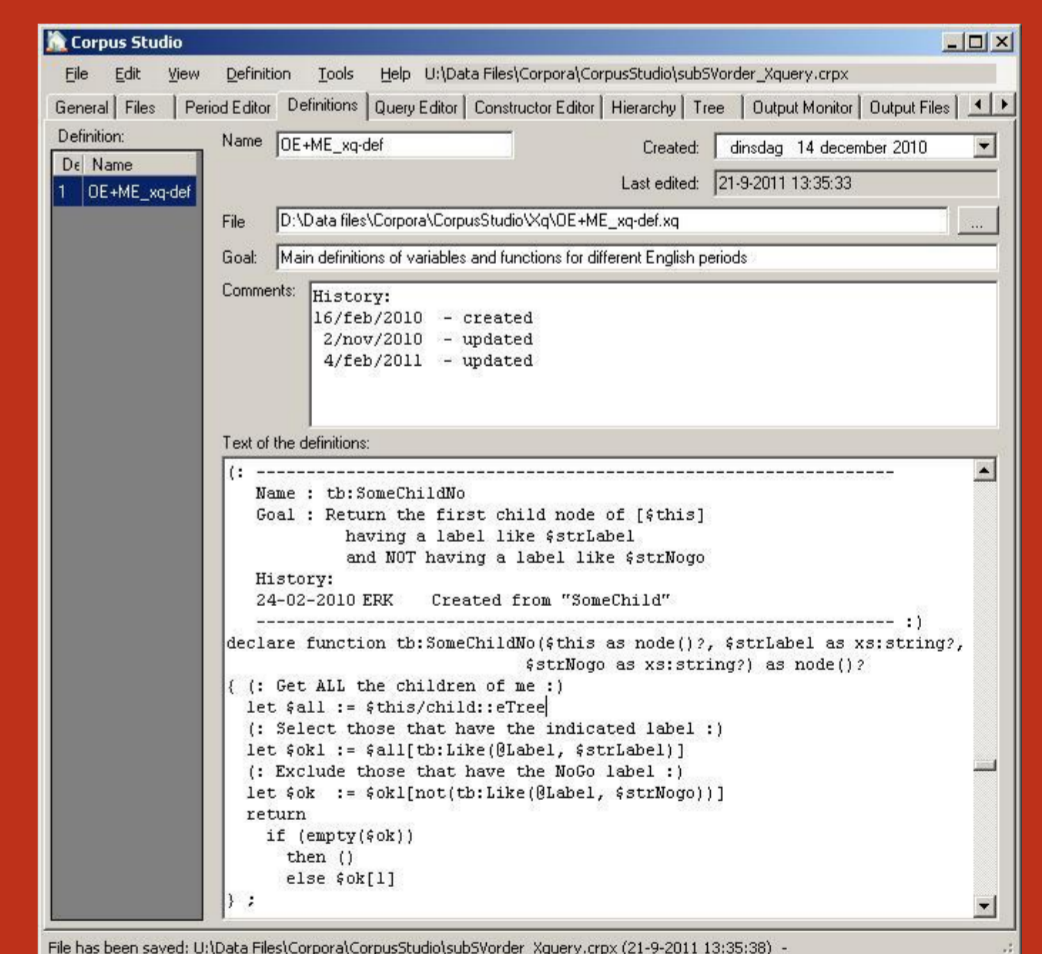
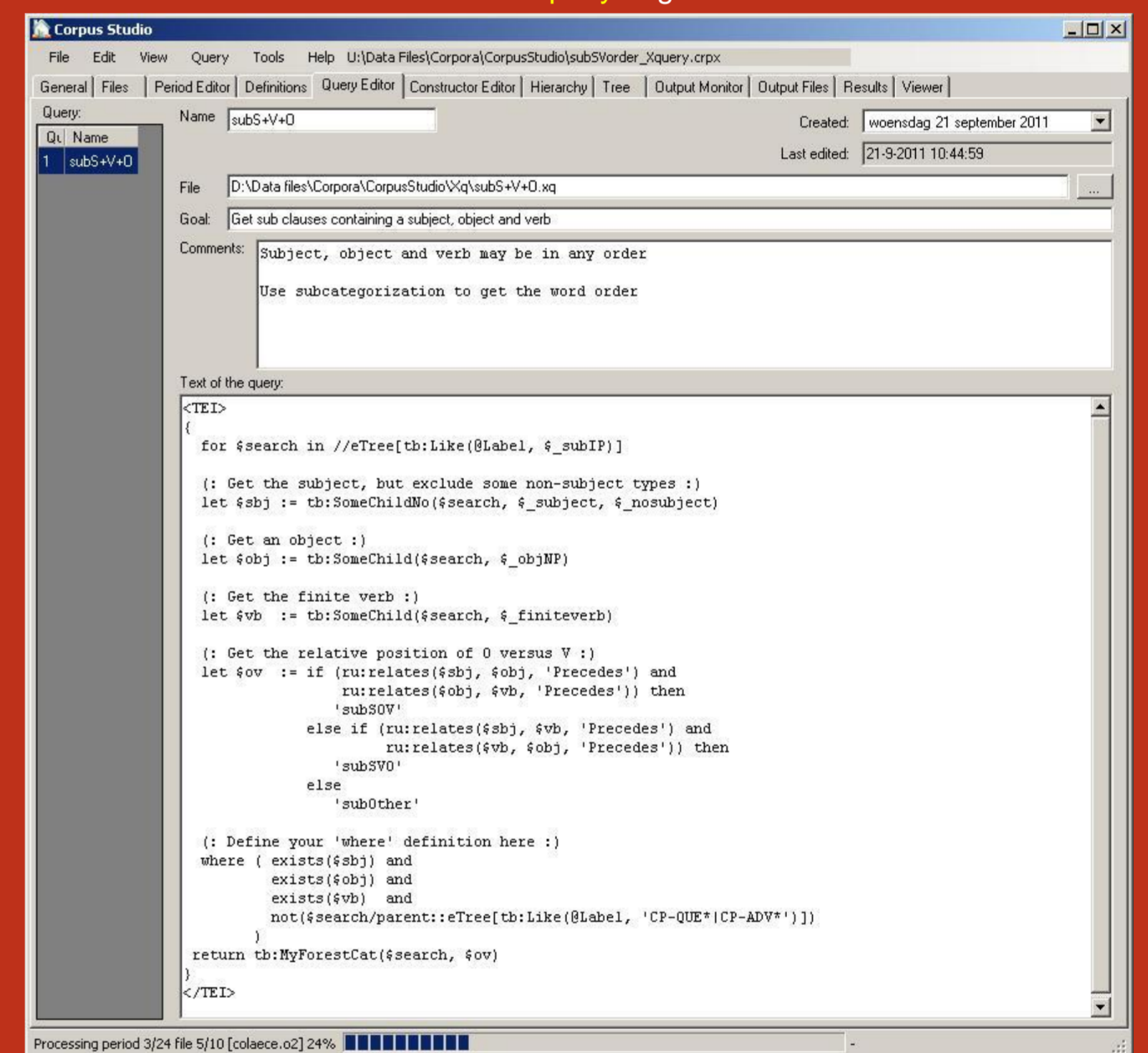
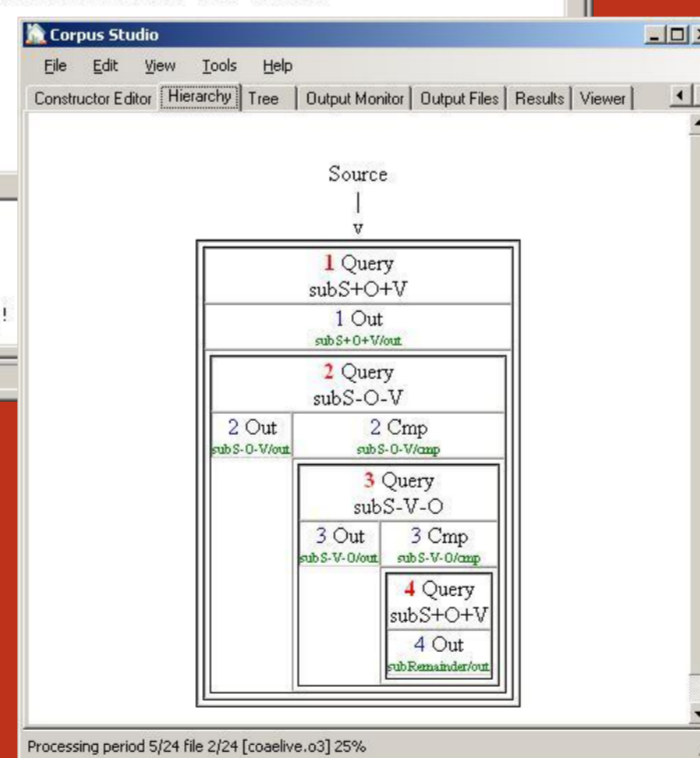
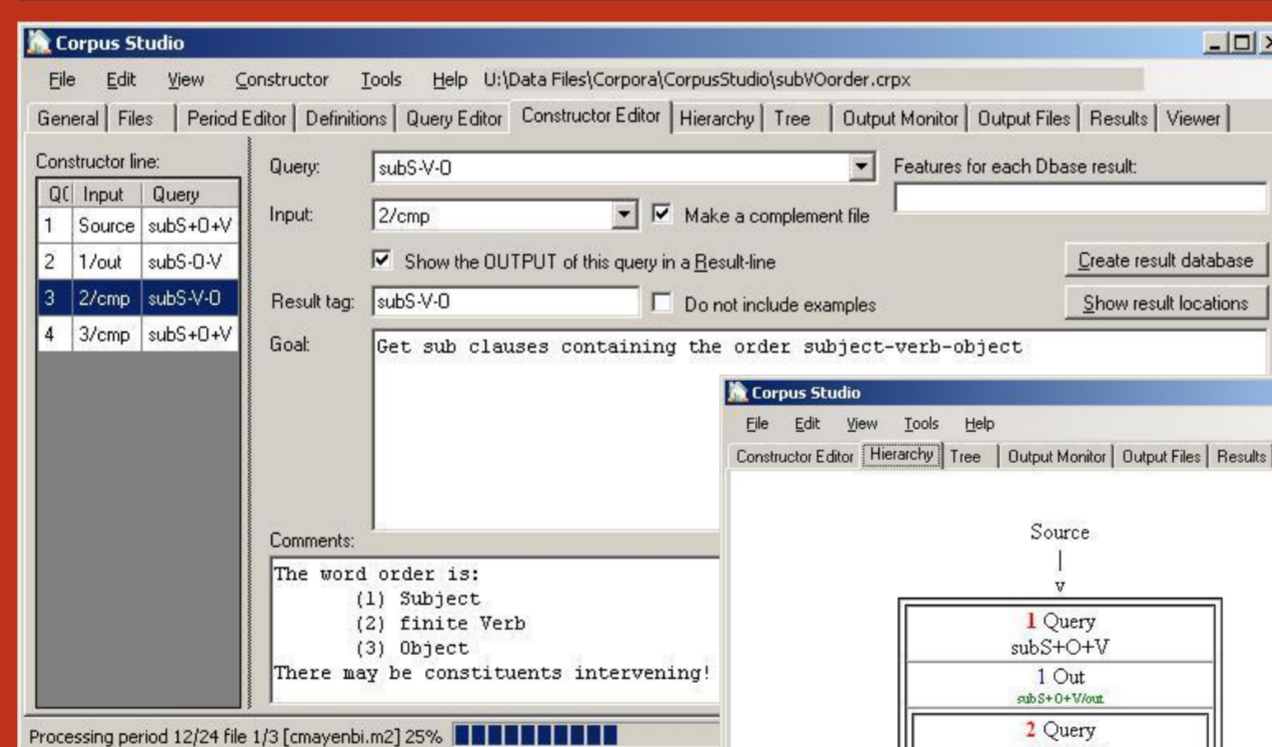
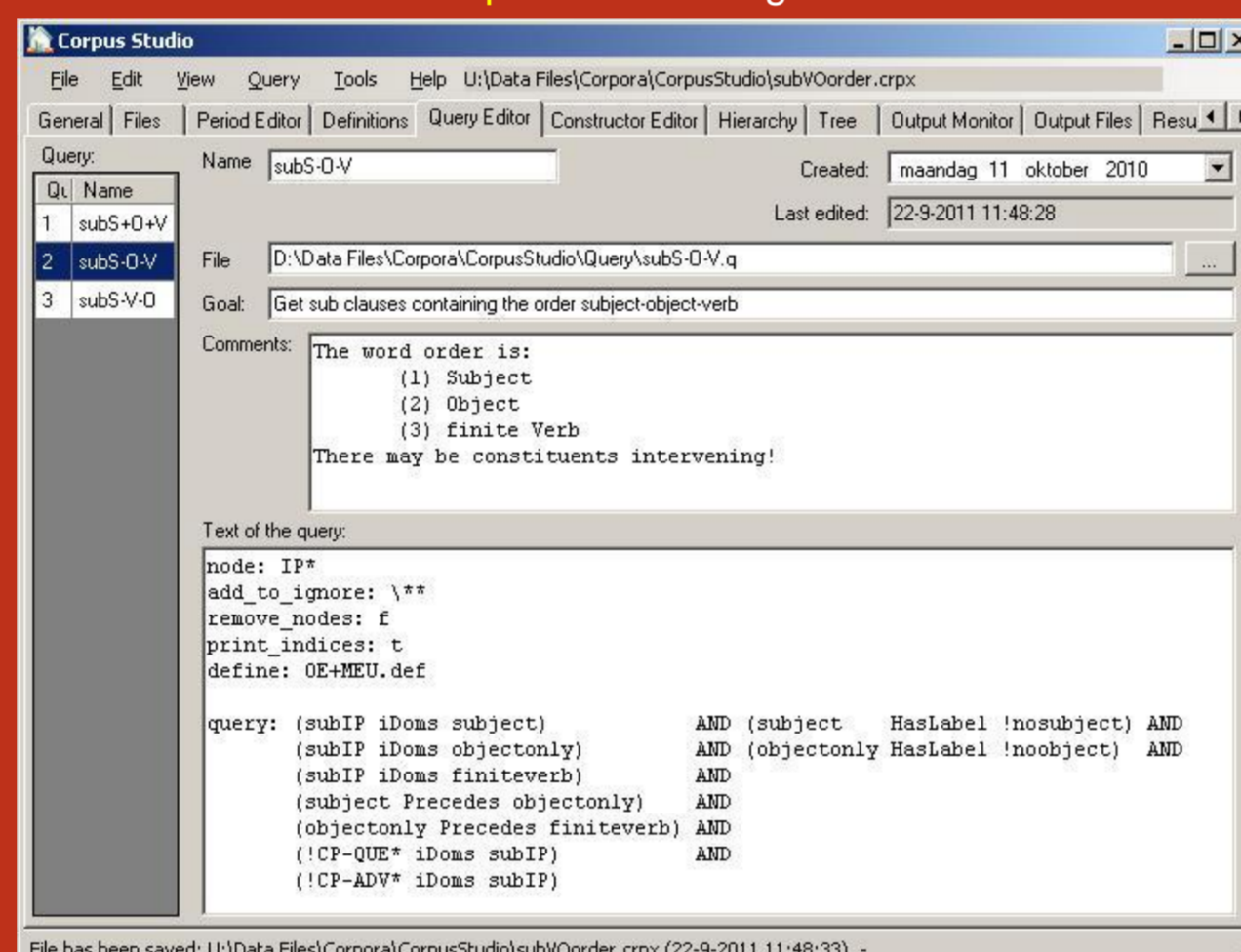
Xquery-psdx project
Xquery engine

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)



Description	O1	O2	O3	O4	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	E1	E2	E3	B1	B2	B3	File				
subS+V+O	17	5	2655	790	3270	1622	813	480	17	1637	447	482	111	275	2575	814	1591	6	3807	4327	2717	2013	1774	1419	subS+V+O
subS+V+O_subOther	4	2	425	108	243	197	110	41	2	87	18	20	2	11	52	18	25	0	72	56	39	37	19	15	(subcategory)
subS+V+O_subSOV	8	2	1462	420	1727	935	313	260	3	330	143	89	0	14	3	2	0	7	8	0	0	0	0	0	(subcategory)
subS+V+O_subSVO	5	1	768	262	1300	490	390	179	5	1220	286	373	109	264	2509	793	1563	6	3728	4263	2678	1976	1755	1404	(subcategory)
IP-MAT	83	13	20315	10786	50201	21428	5986	7448	365	15964	4428	7347	619	1909	26318	6187	19839	88	28194	34614	24944	15424	20326	17201	(CorpusStudio)
IP-STUB	112	12	23857	7486	37879	16668	7196	5501	181	18309	4460	5122	1217	2590	23678	7201	12915	120	32887	35501	33012	17980	17321	13158	(CorpusStudio)

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 Leuzius, 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>>.

