Linguistic resources INL

History

3 online accessible corpora of
- 5 million words
- 27 million words (newspaper)
- 38 million words

Characteristics
- originating from books, newspapers, magazines
- Word, WordPerfect, typesetting
- lemma
- PoS tagged
- no metadata
- bibliografic data

No web interface but Telnet

Used in linguistic research and education

Present situation (1)

Parole Corpus
- 20 million words
- TEI tagged (<p>-level)
- web based interface

For all corpora:
- resources are subject to copyright restrictions
- access: userid and password
- all queries stored in log files

Present situation (2)

TST-centre

Corpus of Spoken Dutch (CGN)

Dutch Spelling Guide
INL Corpora
RBN
CLVV (translation lexica)
etc.

Most data have no metadata (except CGN)

IMDI portal foreseen this year

Infrastructure

Unix/Linux/Windows servers, approx. 700 Gb storage capacity

IT Department
2 system managers
4 system developers

TST-Centre
1 system manager
2 computerlinguists