



Språkbanken · Bank of Swedish
<<http://spraakbanken.gu.se>>

- STpN · Bergen · 25 October 2002
- Lars Borin, Martin Gellerstam,
Jerker Järborg, Rudolf Rydstedt



whence Språkbanken?

- 1975: Logoteket (Sture Allén):
national repository of machine-
readable text (mission: "collect,
process and store electronic texts
for linguistic use")
- first nationally financed, now
through Faculty of Humanities, GU

whence Språkbanken?

- provider of concordances on paper, then microfiche
- materials available electronically from mid-80's, and since early 90's on the Web
- lexicological research and lexicographic work constant driving forces

whence Språkbanken?

- Språkbanken fairly typical: standard corpus \Rightarrow lexicology-centeredness
- ... but comparatively early
- now also language technology concerns
- ... but from a "mature" perspective (and still with lexical focus!)

resources in Språkbanken

- corpora (non-annotated and annotated) (~ 130 million wds)
- lexical resources

corpus resources in Språkbanken

- modern press (1968-), 40 mwd
- modern fiction (1976-), 20 mwd
- historical texts, 9 mwd
- parliament proc., 5.5 mwd

corpus resources in Språkbanken

- PAROLE corpus (POS-annotated), 25 mwd
- SUC corpus (POS, lemmas), 1 mwd
- SALT (parallel, aligned corpora)
- (some syntactically and semantically annotated materials)

lexical resources in Språkbanken

- SAOB (diachronic dictionary of Swedish; 1893-?)
- LEXIN ~ TERMIN
- frequency dictionaries
- lexicology project materials
- (GLDB)

Språkbanken's mission

- collect text materials
- make materials and derived information available to researchers
- build and maintain a "word bank"
- act as project deliverable repository

Språkbanken's mission

- support linguistic research
- information dissemination (to research community and public)
- long-term lexicological/lexicographic work
- (commercial activities)

Språkbanken provides ...

- **resource search functions:**
 - word use (synchronic, diachronic)
 - (diachronic) dictionary lookup
 - grammatical (POS) investigations
 - parallel text search
 - statistics generation
- **consultation services**

whither Språkbanken?

- **new corpus processing/ storage/ access system (DUGA)**
- **virtual corpora for searches**
- **unified storage/annotation format, with standoff annotation (XCES)**
- **user-added annotations**

whither Språkbanken?

- all (modern) Swedish corpora to be POS-tagged and lemmatized
- additional analyses: name(d entitie)s, collocations

whither Språkbanken?

- focused addition of new materials
 - domain-specific non-fiction (popular technical/ science text)
- role *vis-à-vis* new official languages (Bank of Swedish?)
- national mission emphasized by committee report *Mål i mun*

whither Språkbanken?

- more emphasis on resources for LT (as opposed to lexicography)
[still lexicography at the roots!]
- infrastructure and continuity are not sexy, but very important!