GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

CLT

Diabase: Towards a diachronic BLARK in support of historical studies

Lars Borin, Markus Forsberg, Dimitrios Kokkinakis

Språkbanken • Centre for Language Technology
University of Gothenburg

LREC 2010, Valletta, 19th May, 2010

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

topics

1. text and speech as historical research data, and language technology
2. our work with LT for historical studies: two examples
3. methodological musings on BLARKs and language variation

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

LT in historical studies

In historical studies, text – and speech, i.e., language – are central as both primary and secondary research data sources.

In today's world, the normal mode of access to text, speech, images and video is in digital form. Modern material is born digital and older material is being digitized on a vast scale in cultural heritage and digital library projects.

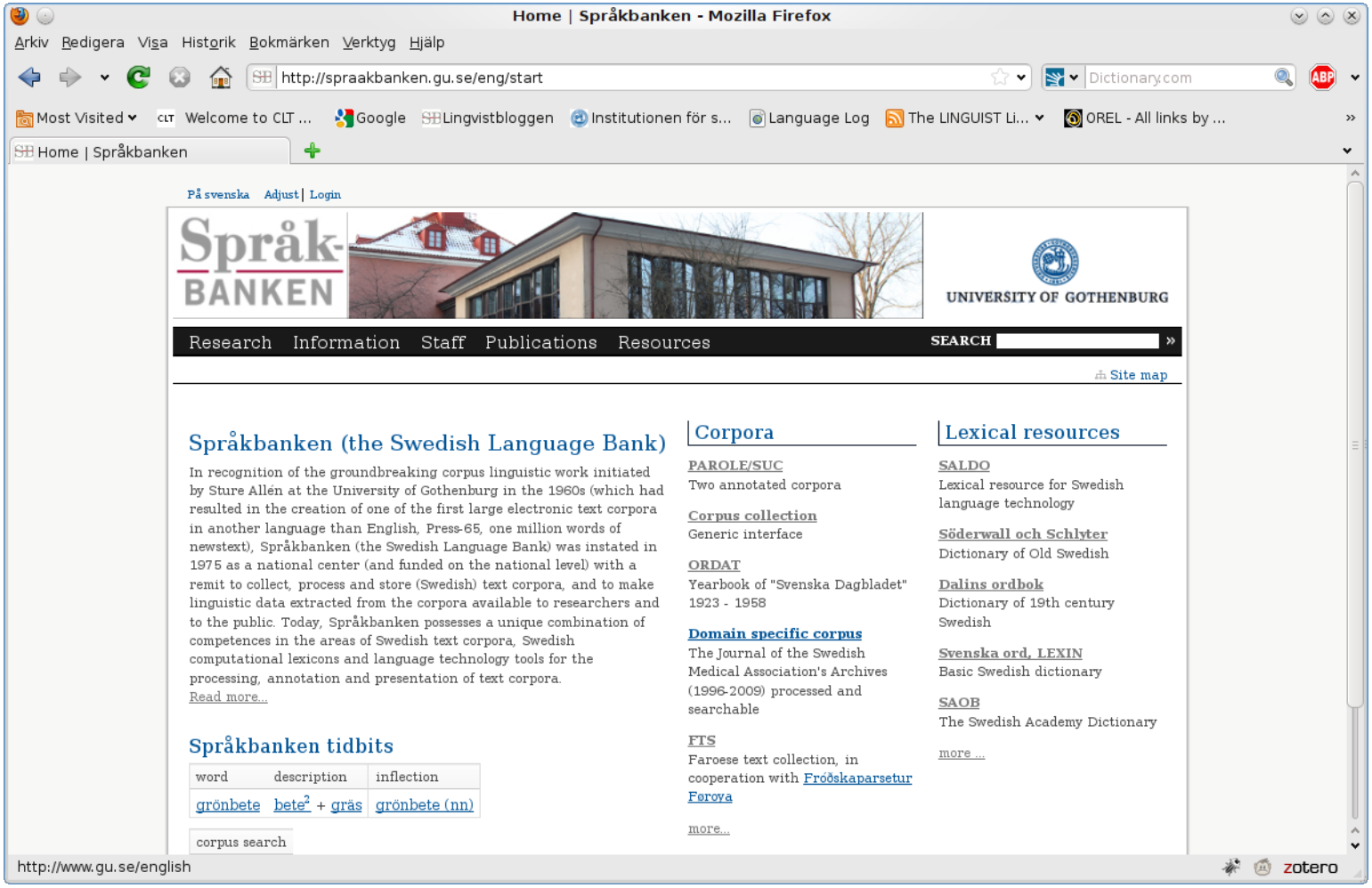
LT can help historians and other researchers make effective use of this flood of language data from all historical periods.

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

http://spraakbanken.gu.se/eng/start/

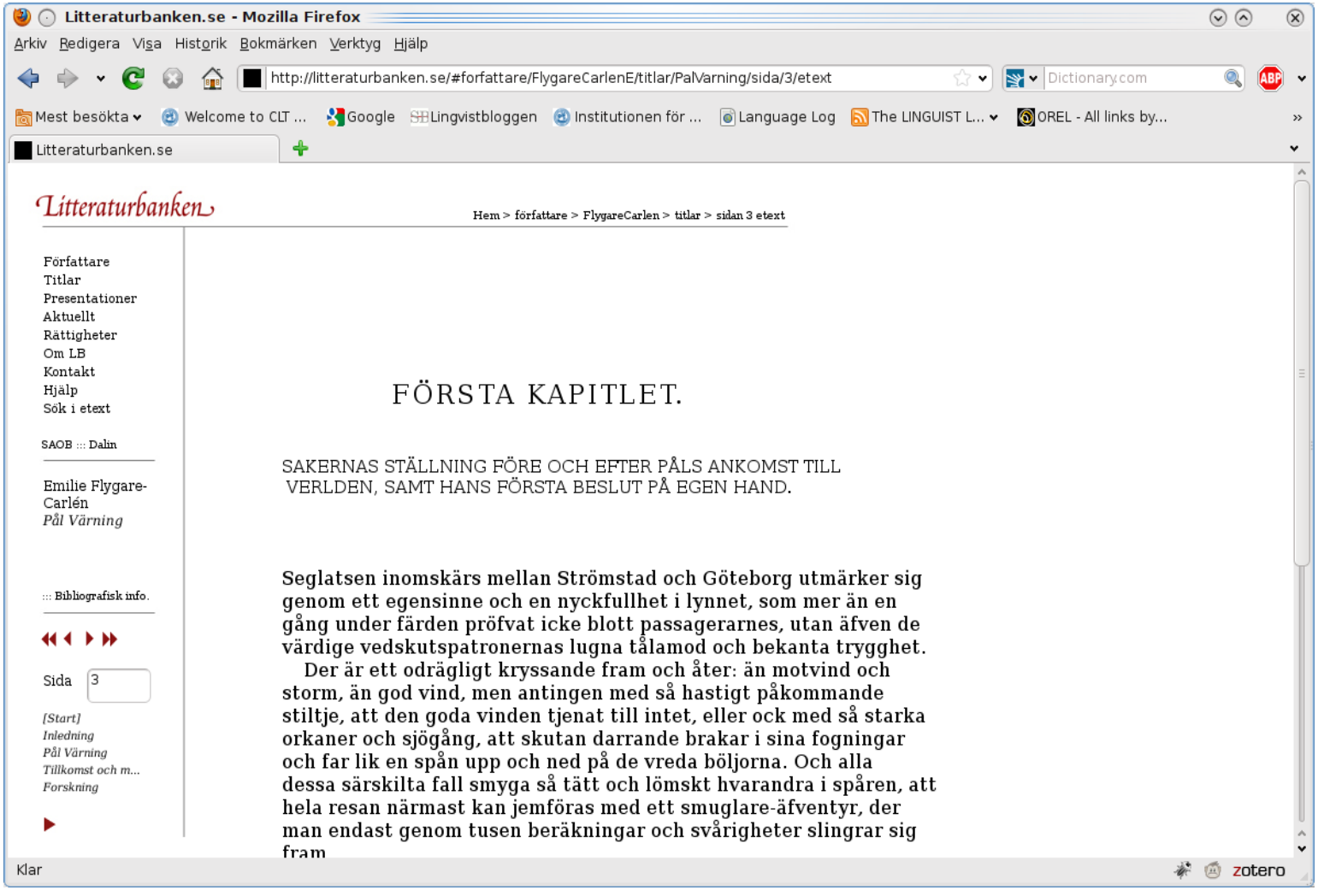


19th c. fiction in Litteraturbanken

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT



GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

NER in Litteraturbanken

Alla författare | Samtliga titlar | Visa 20 träffar per sida

Visar **1021-1040** av **43126** träffa: | Ny sökning | efter förekomst

Platsnamn | Samtliga undertyper

Berger Bendel & Co	s. 284	ig och passerades. Bortom **Adams street** street blef den oerhörda avenyen
Berger Bendel & Co	s. 284	street blef den oerhörda **avenyen** mörkare, och vid van Burens hörn
Berger Bendel & Co	s. 284	en loge. - Kommo de till **Paris** i sommar så skulle de köpa ett par
Berger Bendel & Co	s. 284	hon visste en adress vid **Boulevard Males-** Malesherbes, där de såldes i
Berger Bendel & Co	s. 284	elge någonsin har varit i **Paris**, tillade hon. Van Buren var mörk och
Berger Bendel & Co	s. 284	r som - ligt. Från Lister **Building** hade Helge sett några skepnader
Berger Bendel & Co	s. 285	igt och lyfte lätt på sin **panama**, det är som vore vi i gay Paree –
Berger Bendel & Co	s. 286	Reuter, damerna äro från **Norge** eller, nej, hur var det? Sverige,
Berger Bendel & Co	s. 286	eller, nej, hur var det? **Sverige**, tror jag. Men, herre gud, det är
Berger Bendel & Co	s. 291	nda ned till barriären åt **Clark street** street och bekväma liggstolar, s
Berger Bendel & Co	s. 292	rande, svindlande höjder. **Michigansjöns** ofantliga yta mötte österut
Berger Bendel & Co	s. 293	Snedt till vänster syntes **Masonic Temples** Temples takkrön som ett stråk
Berger Bendel & Co	s. 293	hvilken var reflexen från **Milwaukee.** Men själfva sjön tedde sig som
Berger Bendel & Co	s. 294	kajmurar, hörde han också **Michigans** brusande andhämtning och därefter
Berger Bendel & Co	s. 296	och se efter - gossarna i **England.** Men några dagar hinner jag alltid
Berger Bendel & Co	s. 296	g öfver Kanalen. Bor du i **Paris** på Grand Louvre som förut? - Ja, svarad
Berger Bendel & Co	s. 296	analen. Bor du i Paris på **Grand Louvre** Louvre som förut? - Ja, svarade
Berger Bendel & Co	s. 298	hvars syster är gift med **Indiens** vicekonung. Jag måste gå. Han
Berger Bendel & Co	s. 298	talade de om den franska **rivieran** och spelbanken i Monte Carlo. -
Berger Bendel & Co	s. 298	rivieran och spelbanken i **Monte Carlo** Carlo. - Jag skall hvila ut där i

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

semantic search in 19th c. fiction

CONPLISIT – the components:

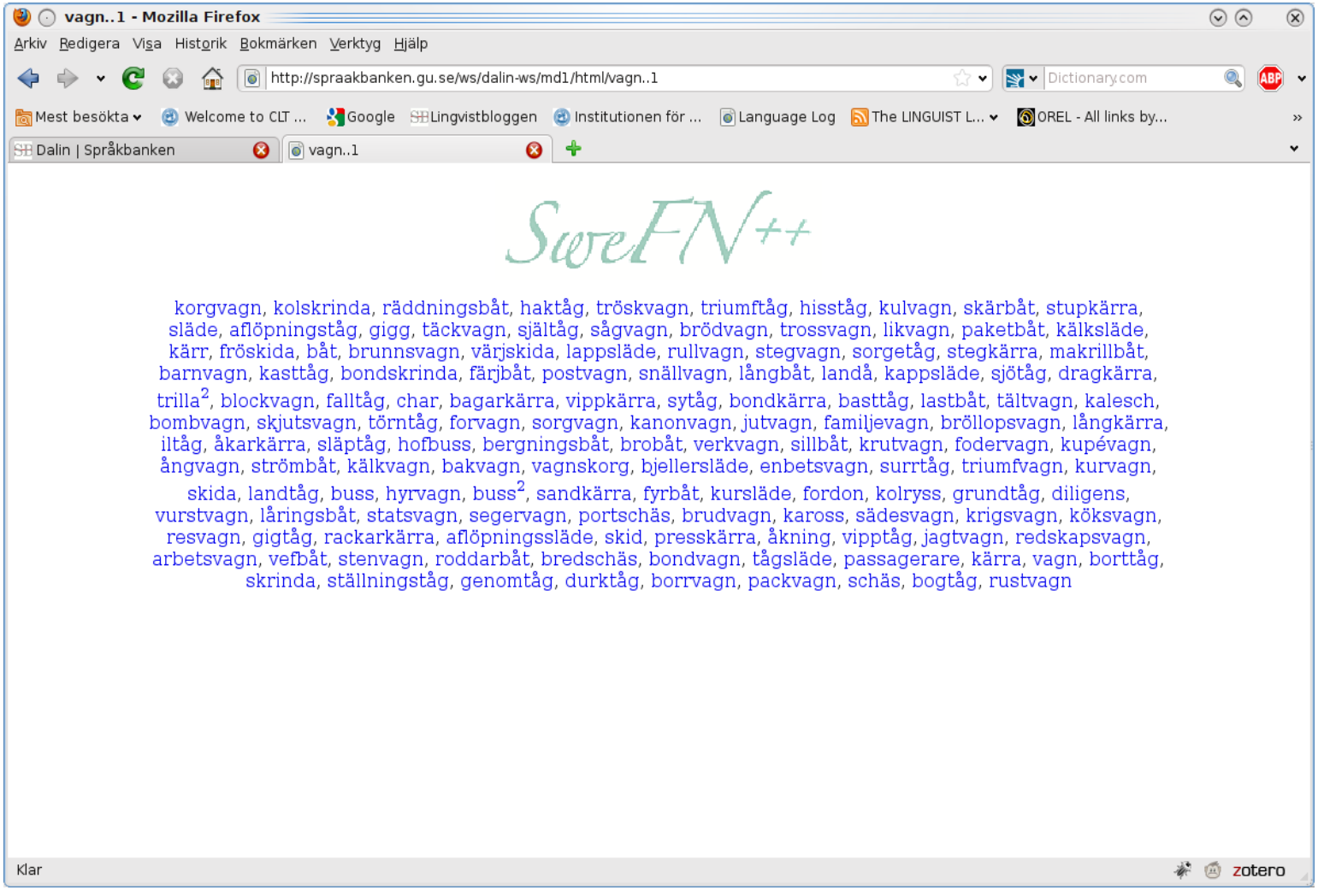
- SALDO – a modern semantic lexicon with inflectional morphology (∼73.000 senses)
- Dalin – a large 19th century lexicon (∼63.000 lemmas)
- an orthographic mapping database SALDO–Dalin
- a morphology for regular 19th c. open parts of speech
- Litteraturbanken (∼100 19th c. novels)
- a research question: can 19th c. fiction throw light on the emergence of consumer society in Sweden?

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

Dalin–SALDO round trip



GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

some issues encountered

- the 1906 spelling reform – in reality lasting some decades around 1900
- large synchronic orthographic variation before the 19th century – and again today!
- slightly different inflectional morphologies 19th–20th century – gradual abandonment of verb-subject agreement during the first half of the 20th century until it officially went out of use around 1950
- very different inflectional system (and syntax) in the Old Swedish period
- changes in vocabulary
- changes in word meanings

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

methodological musings: what's in a BLARK?

- linguistically annotated text corpora
- speech databases
- tools for basic text and speech processing
- basic lexical resources
- tools for lingustic annotation of text (POS taggers, chunkers, parsers)
- text-to-speech and speech-to-text systems

. . . in interoperable, standardized formats

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

BLARKs are static, . . .

As it is normally conceived of and presented, the BLARK assumes a modern standard language variety as the object of description, at least as far as the written language part of the BLARK is concerned, which is the part that we are competent to make judgements about. Part of the reason for this is certainly historical: The BLARK has been – and continues to be – informed more than anything else by language technology work on modern stable written standard languages.

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

. . . language is dynamic

Modern linguistics increasingly recognizes variation as a fundamental and essential characteristic of human language. In this regard, the study of history through textual primary sources makes up an interesting and challenging testbed, where the robustness and the generality of existing language technology are subjected to the acid test of messy and multilingual reality, more so than in many other application areas, since we have to deal with, *inter alia*, historical, non-standardized language varieties in addition to a number of modern standard languages.

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

linguistic variation

Language varies (at least),

- by community (languages, dialects, sociolects)
- by subject, purpose or medium (topics, genres)
- by time (historical language stages)

The BLARK attempts to abstract away from all three; it can be thought of as reflecting a modern standard language, which is topic- and genre-neutral.

GÖTEBORGS UNIVERSITET

Språk-BANKEN

CLT

describing variation in a BLARK?

Our work described above can be seen as the first steps towards the development of *Diabase*, a Swedish BLARK extended along the diachronic – or time – axis. This work also raises some interesting methodological questions about the description of variation in the linguistic resources making up the BLARK:

- fundamental/norm(al) $\sim$ deviation (e.g., 'correct' vs. 'incorrect' verb agreement in a novel published in 1935; correct vs. incorrect spelling in an internet page in 2010)
- one system $\sim$ a mix of systems (e.g., pre- vs. post-1906 spelling)

GÖTEBORGS
UNIVERSITET

Språk-
BANKEN

CLT

thank you for listening!