

GÖTEBORGS UNIVERSITET



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

◆□▶ ◆□▶ ▲□▶ ▲□▶ □ のQで



topics

GÖTEBORGS UNIVERSITET

> Språk-BANKEN

CLT

1. text and speech as historical research data, and language technology

◆□▶ ◆□▶ ▲□▶ ▲□▶ □ のQ@

- 2. our work with LT for historical studies: two examples
- 3. methodological musings on BLARKs and language variation



LT in historical studies

GÖTEBORGS UNIVERSITET



CIT

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.

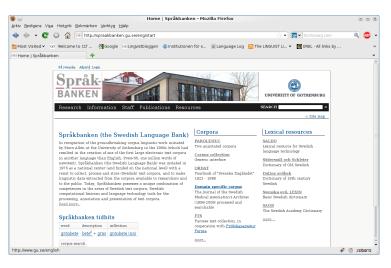


http://spraakbanken.gu.se/eng/start/





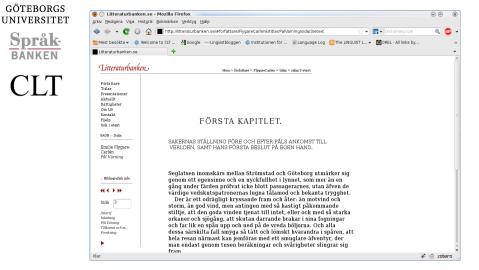
CLT



<ロト < 同ト < 回ト < 回ト = 三日



19th c. fiction in Litteraturbanken





NER in Litteraturbanken



Alla författare 🔽	Sam	tliga titlar		Visa 20 🗸 träffar per sida
Visar 1021-1040 av 43126 träffa 🖂 🔷 🗁 🖂			Ny sökning	efter förekomst 🔹
Platsnamn 🔽 Samtliga undertype		a undertyper		
Denner Dendel & ener Or	- 004	in a b a com	des Dertem Adams stores	atorest blaf dan ande inde anonen
Berger Bendel & amp; Co		ig och passerades. Bortom Adams street street blef den oerhörda avenyen		
Berger Bendel & amp; Co		street blef den oerhörda avenyen mörkare, och vid van Burens hörn		
Berger Bendel & amp; Co		en loge Kommo de till Paris i sommar så skulle de köpa ett par		
Berger Bendel & amp; Co		hon visste en adress vid Boulevard Males- Malesherbes, där de såldes i		
Berger Bendel & amp; Co	s. 284	elge någonsin har varit i Paris, tillade hon. Van Buren var mörk och		
Berger Bendel & amp; Co	s. 284	r som - ligt. Från Lister Building hade Helge sett några skepnader		
Berger Bendel & amp; Co	s. 285	igt och lyfte lätt på sin panama, det är som vore vi i gay Paree -		
Berger Bendel & amp; Co	s. 286	Reuter, damerna äro från Norge eller, nej, hur var det? Sverige,		
Berger Bendel & amp; Co	s. 286	eller, nej, hur var det? Sverige, tror jag. Men, herre gud, det är		
Berger Bendel & amp; Co	s. 291	nda ned till barriären åt Clark street street och bekväma liggstolar, s		
Berger Bendel & amp; Co	s. 292	rande, svindlande höjder. Michigansjöns ofantliga yta mötte österut		
Berger Bendel & amp; Co	s. 293	Snedt till vänster syntes Masonic Temples Temples takkrön som ett strål		
Berger Bendel & amp; Co	s. 293	hvilken var reflexen från Milwaukee. Men själfva sjön tedde sig som		
Berger Bendel & amp; Co	s. 294	kajmurar, hörde han också Michigans brusande andhämtning och därefter		
Berger Bendel & amp; Co	s. 296	och se efter - gossarna i England. Men några dagar hinner jag alltid		
Berger Bendel & amp; Co	s. 296	g öfver Kana	llen. Bor du i Paris på Gran	d Louvre som förut? – Ja, svarad
Berger Bendel & amp; Co	s. 296	analen. Bor du i Paris på Grand Louvre Louvre som förut? – Ja, svarade		
Berger Bendel & amp; Co	s. 298	hvars syste	r är gift med Indiens viceko	nung. Jag måste gå. Han
Berger Bendel & amp; Co	s. 298	talade de om den franska rivieran och spelbanken i Monte Carlo		
Berger Bendel & amp; Co	s. 298	rivieran och	spelbanken i Monte Carlo	Carlo. – Jag skall hvila ut där i



semantic search in 19th c. fiction

GÖTEBORGS UNIVERSITET



CLT

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?

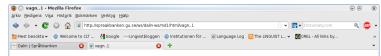


Dalin-SALDO round trip





CIT



SureFN++

korywan, kolskrinda, riddningsbåt, haktåg, tröskvagn, triumfåg, hisståg, kulvagn, skärbåt, stupkära, slåda, afönningstå, gigg, täkvagn, sjätlag, sövagn, brövdvagn, trossvagn, likvagn, paketbåt, kälkislada, kärr, fröskida, båt, brunnsvagn, värjskida, lappeläde, rullvagn, stegvagn, sorgetåg, stegkära, makrillbåt, barrvagn, kaståd, pondskrinda, färbåt, postvagn, snällvagn, lindbåt, landå, kapsislda, gölföd, drackära,

trilla², blockvagn, faildag, char, bagaräärra, vippkärra, sytäg, bondkärra, bastkåg, lastbåt, tältvagn, kalesch, bonbvagn, skultvagn, törindagn, förvagn, sorgvagn, kanorvagn, jutvagn, familjövagn, bröllöpsegn, längkärs, litäg, kärrkärra, släptåg, hofbuss, bergningsbåt, brobåt, verkvagn, silibåt, krutvagn, fodervagn, kupdvagn, angvagn, strömbåt, kältvagn, bakvagn, vagnskorg, bjellersläde, enbetsvagn, surtdåg, trumprägn, kurvagn, kupdvagn, kurvagn, kupdvagn, kurvagn, kupdvagn, kurvagn, kurvagn, kupdvagn, kurvagn, k

skida, landtåg, buss, hyrvagn, buss², sandkärra, fyrbåt, kurslåde, fordon, kolryss, grundtåg, diligens, vurstvagn, läringsbåt, statsvagn, segervagn, portschäs, brudvagn, kranss, sådesvagn, krigsvagn, köksvagn, resvagn, gitda, rackarkärra, aföpningssläde, skid, presskara, äkning, vippvlad, jadyvagn, redskapsvagn, arbetsvagn, vefbåt, stenvagn, roddarbåt, bredschäs, bondvagn, tågsläde, passagerare, kärra, vagn, borttåg, skrinda, ställningstäd, genomtåg, durktåd, bortvagn, pastvagn, schäs, bogtåg, rustvagn

atero 🖉 🧩

▲□▶ ▲□▶ ▲□▶ ▲□▶ □ のQで

Klar



GÖTEBORGS UNIVERSITET



CLT

some issues encountered

- the 1906 spelling reform in reality lasting some decades around 1900
- Iarge 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



methodological musings: what's in a BLARK?



UNIVERSITET

CLT

- 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

◆□▶ ◆□▶ ▲□▶ ▲□▶ □ のQで

... in interoperable, standardized formats



BLARKs are static, ...

GÖTEBORGS UNIVERSITET

> Språk-BANKEN



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.



UNIVERSITET

Språk-

RANKEN

CIT

... 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.



linguistic variation

GÖTEBORGS UNIVERSITET



CIT

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.



describing variation in a BLARK?

GÖTEBORGS UNIVERSITET



CLT

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) ~ 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 ~ a mix of systems (e.g., pre- vs. post-1906 spelling)



thank you for listening!

GÖTEBORGS UNIVERSITET



CLT

