Historical Corpora

Yvonne Adesam
Språkbanken
University of Gothenburg

2013

Outline

Corpus usage

Historical corpus linguistics

Examples

Lexical link-up

Why historical corpora?

Questions:

- How did *do* support develop in English?
- How has the meaning of *experimental* changed?
- How is the emerging consumer society depicted in literature?

English *do*-support

Auxiliary *do* used in yes-no questions and negative sentences in English:

a) Some people (*do) love honour and truth

b) Some people *love not / do not love honour and truth

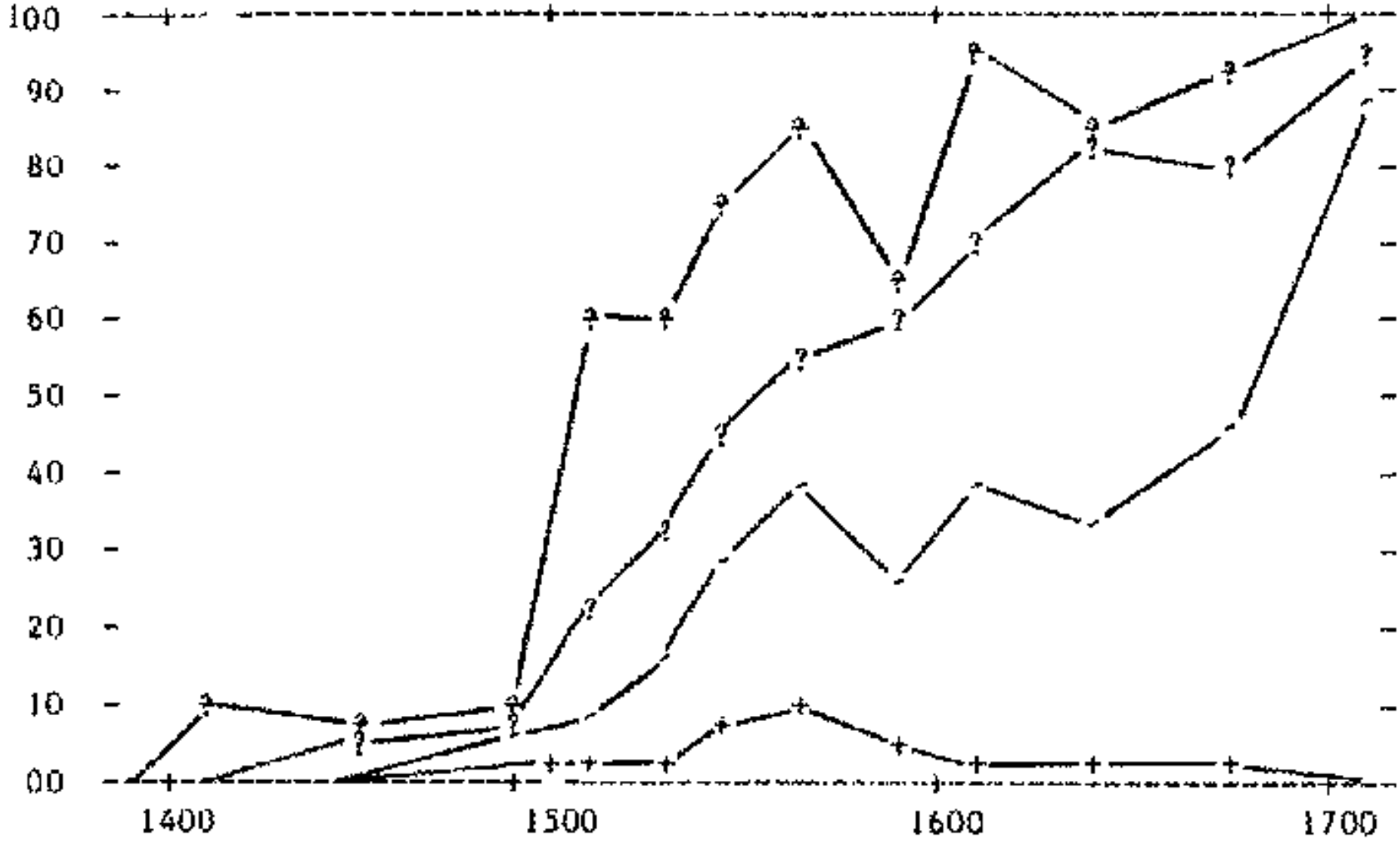
c) *Love some people / Do some people love honour and truth?

d) Is ther no morsel breed that ye do keepe?
'Isn't there a bit of bread that you've saved?'

e) Fly fro company of them that lovyth not honour and throuthe.
'Get away from those who do not love honour and truth.'

[Ellegård, 1953] collected 10k sentences between 1400–1700 with/without 'do' to study rise. Reanalysis in [Kroch, 1989].

English *do*-support



Kroch (1989), data from Ellegård

The change of *experimental*

How has the meaning of *experimental* changed?
[Pumfrey et al., 2012]

- *experiment*: religious → scientific in 1660s
- Large collection of Early English Books
- Manual exploration
  - *experiment* with variants returned 21k hits in 3k records
  - Manual copy-and-paste and inspection
  - Manual classification and count (frequency normalization)
  - 4 weeks work
- "Automated methods can achieve similar results in minutes"
  - Explore data as concordance (fast through pre-indexing)
  - Frequency by decade, possibly automated annotation
  - Fast retrieval, automatic counting
  - Permits iterative hypothesis testing, data driven investigation

Consumption patterns and life-style in literature

How is the emerging consumer society depicted in Swedish literature?
[Borin et al., 2011]

- Using available resources
  - Modern semantic lexicon and morphology
  - 19th c. lexicon and morphology
- Algorithm for semantic search
  - Look up text word in the morphologies
  - Collect all associated senses in the semantic lexicon
  - List all entries in historical lexicon connected to these senses

Consumption patterns and life-style in literature

Figure 2: Word form lookup of *soffa* 'sofa'

[Borin et al., 2011]

Consumption patterns and life-style in literature

uppmöblera	kanapé	bårstol	valkbord	tryckbord	tingsbord	tebord	svärtbord	strykbord	stigbord
spegelbord	skådebord	skärbord	skänkbord	ljusbord	kredensbord	kammarbord	hästskobord	herrskapsbord	färgbord
fällbord	friserbord	fortunabord	formbord	dambord	bröstbord	brädspelsbord	bord	bakbord	altarbord
tidningsbyrå	ottoman	kommissionsbyrå	klädesbyrå	divan	byrå	bord[2]	affärsbyrå	adressbyrå	vändstol
väggbänk	vridstol	verkstol	verkbänk	vaskbänk	varpstol	varmbänk	valsstol	vaktbänk	understol
tvättstol	tvärbänk	torfbänk	sågbänk	svarfbänk	strumpstol	stol	sqvalbänk	spegelbänk	sofstol
soffstol	slipbänk	slagtbänk	slagtarbänk	skärbänk	skottstol	skjutstol	rörstol	ryggstol	rullbänk
rottingstol	reffelbänk	pressbänk	plantbänk	pinbänk	likstol	liggstol	kullerstol	korbänk	klädesstol
kardbänk	kalkbänk	häftstol	hvilostol	huggbänk	hjulstol	handstol	halmstol	gubbstol	gräsbänk
fogbänk	fallbänk	erkebiskopsstol	dufstol	dragstol	bönstol	bänk	brudbänk	borrstol	borrbänk
bordbänk	bokstol	biskopsstol	bibänk	bandstol	armstol	vaksäng	trädgårdssäng	säng	syskonsäng
sparrissäng	skogsäng	sjuksäng	plantsäng	paulunsäng	negliksäng	melonsäng	löksäng	kryddsäng	korgsäng
gurksäng	fältsäng	dödssäng	bröllopssäng	blomsäng	blomstersäng	bergsäng	utdragssoffa	soffa	hvilosoffa
skeppsbord									

Figure 3: Words semantically related to *soffa* 'sofa'

[Borin et al., 2011]

Historical corpora

Just like other corpora

- allow for looking through large quantities of data
- allow for querying large text collections over and over

Maybe especially important for historical data

- difficult to access
- hard to read

The pitfalls of historical corpus linguistics

3 problems for using historical corpora [Rissanen, 1989], see also [Rissanen, 2008].

- The pedagogical "philologist's dilemma"
- The methodological "God's truth fallacy"
- The pragmatic "mystery of vanishing reliability"

Not reasons against corpus linguistics, but warnings to take care when building and using (historical) corpora

The pitfalls of historical corpus linguistics

3 problems for using historical corpora [Rissanen, 1989], see also [Rissanen, 2008].

- The pedagogical "philologist's dilemma"
  Studying the original texts in their contexts gives in-depth knowledge of language history
- The methodological "God's truth fallacy"

- The pragmatic "mystery of vanishing reliability"

Not reasons against corpus linguistics, but warnings to take care when building and using (historical) corpora

The pitfalls of historical corpus linguistics

3 problems for using historical corpora [Rissanen, 1989], see also [Rissanen, 2008].

- The pedagogical "philologist's dilemma"
  Studying the original texts in their contexts gives in-depth knowledge of language history
- The methodological "God's truth fallacy"
  Corpora are limited in what they cover and are never fully representative
- The pragmatic "mystery of vanishing reliability"

Not reasons against corpus linguistics, but warnings to take care when building and using (historical) corpora

The pitfalls of historical corpus linguistics

3 problems for using historical corpora [Rissanen, 1989], see also [Rissanen, 2008].

- The pedagogical "philologist's dilemma"
  Studying the original texts in their contexts gives in-depth knowledge of language history
- The methodological "God's truth fallacy"
  Corpora are limited in what they cover and are never fully representative
- The pragmatic "mystery of vanishing reliability"
  More variables (periods, genres, age, gender etc) means less statistical reliability

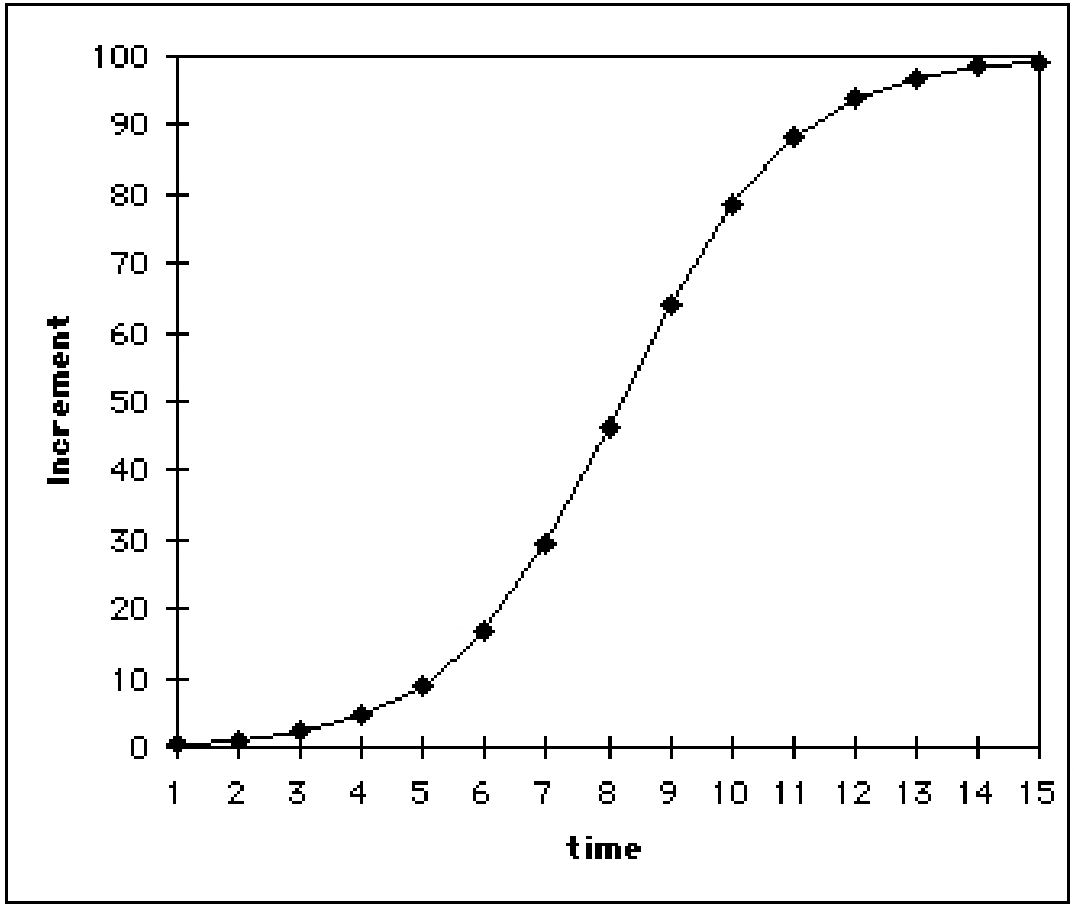
Not reasons against corpus linguistics, but warnings to take care when building and using (historical) corpora

Language change

- Living languages have rich variation, seed for language change
- Any living language is in constant change
- Extra-linguistic factors for change
  - Sociolinguistic (status, education etc)
  - Contextual (medium, topic etc)
  - Regional (incl contact)
- Language-internal factors for change
  - Grammaticalization
  - Metaphors
  - Emphatic expressions

Language change, cont.

At any one time two or more competing variants co-exist



Some historical corpora

- Helsinki corpus of English texts (9th–18th c.)
- ARCHER – Historical English Registers (17th–20th c.)
- Zürich Corpus of English Newspapers (17th–18th c.)
- IcePaHK Islandic treebank (12th–21st c.)
- GerManC German representative corpus (17th–18th c.)
- Tüba-D/DC with German Gutenberg material (13th–20th c.)
- http://spraakbanken.gu.se/korp

Old Swedish Text

Explore the Old Swedish text!

How can we convert this into a corpus?
What could be problematic?
What could be easy?

Old Swedish Text

Han beddis almoso af petro ok johanne Tha han saa them til byriä at inga j mönstrit ‖ Petrus sagde til hans ‖ Jak hafuir ey gul ällär silfuer vtan thz som iak hafuir gifuir jak thik ‖ j ihesu christi nazareni nampn stat vp ok gak ‖ Ok ginstan grep sanctus petrus hans höghre hand ok vplypte han ok ämbrat festos hans sinor ok fötir ok han sprang vp stodh ok gik in j mönstrit medhär thöm gangande ok springande ok lowande gudh ‖ ok alt folkit saa han gangande ok lofuande gudh ok kiändo han at han var then sami som saat for mönstersins port thiggiande almoso ok vndradho mykit a thz som honom var hänt ‖ ok then tidh the hioldo petrum oc iohannem lop alt folkit til therä vndrande

Old Swedish Text

Han beddis almoso af petro ok johanne Tha han saa them til byriä at inga j mönstrit || Petrus sagde til hans || Jak hafuir ey gul ällär silfuer vtan thz som iak hafuir gifuir jak thik || j ihesu christi nazareni nampn stat vp ok gak || Ok ginstan grep sanctus petrus hans höghre hand ok vplypte han ok ämbrat festos hans sinor ok fötir ok han sprang vp stodh ok gik in j mönstrit medhär thöm gangande ok springande ok lowande gudh || ok alt folkit saa han gangande ok lofuande gudh ok kiändo han at han var then sami som saat for mönstersins port thiggiande almoso ok vndradho mykit a thz som honom var hänt || ok then tidh the hioldo petrum oc iohannem lop alt folkit til therä vndrande

Problematic issues:

- Sentence boundaries

Old Swedish Text

Han beddis almoso af petro ok johanne Tha han saa them til byriä at inga j mönstrit ‖ Petrus sagde til hans ‖ Jak hafuir ey gul ällär silfuer vtan thz som iak hafuir gifuir jak thik ‖ j ihesu christi nazareni nampn stat vp ok gak ‖ Ok ginstan grep sanctus petrus hans höghre hand ok vplypte han ok ämbrat festos hans sinor ok fötir ok han sprang vp stodh ok gik in j mönstrit medhär thöm gangande ok springande ok lowande gudh ‖ ok alt folkit saa han gangande ok lofuande gudh ok kiändo han at han var then sami som saat for mönstersins port thiggiande almoso ok vndradho mykit a thz som honom var hänt ‖ ok then tidh the hioldo petrum oc iohannem lop alt folkit til therä vndrande

Problematic issues:

- Lack of a standardized orthography

Variation

Lack of standard orthography!
Not just problematic for annotation

- How to count sentences/words
- How to find words

Handling spelling variation

Normalization to a standard orthography.

Pros:

- Quick solution to the type problem
- Robust – unknown words
- Robust – inflection

Cons:

- target normalization may contain irregularities
- no extra information from normalization itself

Approaches:

- rewrite rules
- (Soundex etc)

Handling spelling variation

Link tokens in running text to a lexicon:

Pros:

- Entries as type identifiers
- Access to part-of-speech
- Meaning information (in prose)

Cons:

- Coverage
- Reliance on lexicon
- spelling variation + inflection! (8k / 162k types match in our Old Swedish corpora & lexica)

Approaches:

- fuzzy matching: minimum distance / all within maximum distance
- normalization of corpus *and* lexicon

Levenshtein Distance

The number of operations/character edits – here: insert, delete, substitute – needed to change one string into another.

bokstafwa vs *bogstaffua*

Levenshtein Distance

The number of operations/character edits – here: insert, delete, substitute – needed to change one string into another.

bokstafwa vs *bogstaffua*

bokstafwa $\rightarrow$ bogstafwa 1 substitute

Levenshtein Distance

The number of operations/character edits – here: insert, delete, substitute – needed to change one string into another.

bokstafwa vs *bogstaffua*

bokstafwa	→	bogstafwa	1 substitute
bogstafwa	→	bogstaffwa	1 insert

Levenshtein Distance

The number of operations/character edits – here: insert, delete, substitute – needed to change one string into another.

bokstafwa vs *bogstaffua*

bokstafwa	→	bogstafwa	1 substitute
bogstafwa	→	bogstaffwa	1 insert
bogstaffwa	→	bogstaffua	1 substitute

LD shortcomings: indiscriminative

Text: *tiuku* 'twenty'
Dict: **tiughu.Num**
$d_L = 2$

But also at $d_L = 2$ from *tiuku*:
thuka.N 'darkness, mist'; **tiugh.N** 'collection of twenty'; **tiund.N** '(one) tenth'; **tiuva.V** 'to steal'; **siuke.N** 'disease'; **fiuka.V** 'to blow away'; **miuka.V** 'to soften, be humble, better oneself'; **riuka.V** 'to (produce) smoke'; **tukt.N** 'discipline'; **tik.N** 'bitch'; **tiu.Num** 'ten'

One possible solution

Use weights to distinguish cheap operations (likely correspondences) from expensive (unlikely) ones.

Weights from:

- (informal) corpus inspection, intuition / knowledge of the language
- collecting attested variants (e.g., dictionary), manual rule distillation
- automatic rule extraction from given variants [Adesam et al., 2012]
- automatic collection of variants, etc (bootstrapping / magic!)

Automatically extracted rules

A small sample from 6 045 *n–m* rules [Adesam et al., 2012]:

Rule	Wght	Schlyter example
u→o	0.20	*arvuþi ærvoþi* ‘work’
æ→e	0.27	*ær er* ‘scar’
pt→ft	0.31	*apter after æftær* ‘after’
g#→gg#	0.42	*væg vægg vegg* ‘wall’
þer→n	0.43	*maþer man* ‘man’
au→ö	0.44	*barnlös barnalös barnalaus* ‘childless’
th→þ	0.44	*oþolskipti othalskipte* ‘(type of) land redivision’
mp→m	0.45	*hamn hampn* ‘harbour’
li→eli	0.45	*lastelika lastlika* ‘blameworthy, shameful’
ghi→i	0.62	*aplöia opplöghia* ‘plowing into the neighbour’s field’

Normalization strategies

Rules can have a 'real' target (e.g., modern orthography)...

$h \rightarrow \epsilon \ / \ \# \_ \ v$
$h \rightarrow \epsilon \ / \ g \_$
$k \rightarrow g \ / \ \mathrm{V} \_ \ \mathrm{V}$
$i \rightarrow j \ / \ [t \ k \ s] \_ \ [ä \ y \ u \ ö \ e]$
$u \rightarrow o \ / \ \_ \ \#$
$qu \rightarrow kv$ etc

...or just simplify to remove unimportant or hard distinctions (neutralization).

$[w \ v \ u \ o \ hv \ hu] \rightarrow u$
$[i \ j \ ii \ ij] \rightarrow i$
$[ä \ æ \ e] \rightarrow e$
$[h] \rightarrow \epsilon$ etc.

Summary

- Historical corpora
  - accessible
  - reusable
  - large
- Variation problematic
  - statistics (variation in punctuation, spelling etc)
  - finding words (variation in spelling etc)

References I

Adesam, Y., Ahlberg, M., and Bouma, G. (2012).
bokstaffua, bokstaffwa, bokstafwa, bokstaua, bokstawa. . . towards lexical link-up for a corpus of Old Swedish.
In Jancsary, J., editor, *Empirical Methods in Natural Language Processing: Proceedings of KONVENS 2012 (LThist 2012 workshop)*, pages 365–369, Vienna.

http://www.oegai.at/konvens2012/proceedings.shtml.

Borin, L., Forsberg, M., and Ahlberger, C. (2011).
Semantic search in literature as an e-humanities research tool: Conplisit – consumption patterns and life-style in 19th century swedish literature.
In *NEALT Proceedings Series (NODALIDA 2011 Conference Proceedings)*, volume 11, pages 58–65.
http://hdl.handle.net/10062/17290.

Ellegård, A. (1953).
The Auxiliary do: The Establishment and Regulation of its Use in English.
Almqvist & Wiksell, Stockholm.

References II

Kroch, A. (1989).
Function and grammar in the history of English: periphrastic *do*.
In Fasold, R. and Shiffrin, D., editors, *Language Change and Variation*, volume 52 of *Current Issues in Linguistic Theory*. Benjamins.
ftp://babel.ling.upenn.edu/papers/faculty/tony_kroch/papers/function-grammar-do.pdf.

Pumfrey, S., Rayson, P., and Mariani, J. (2012).
Experiments in 17th century English: manual versus automatic conceptual history.
Literary and Linguistic Computing.
http://llc.oxfordjournals.org/content/early/2012/06/01/llc.fqs017.

Rissanen, M. (1989).
Three problems connected with the use of diachronic corpora.
ICAME Journal, 13:16–19.
http://icame.uib.no/archives/No_13_ICAME_Journal_index.pdf.

Rissanen, M. (2008).
Corpus linguistics and historical linguistics.
In Lüdeling, A. and Kytö, M., editors, *Corpus Linguistics: an International Handbook*, volume I, pages 53–68. Walter de Gruyter, Berlin and New York.