Historical resources

Yvonne Adesam    Gerlof Bouma
Språkbanken
University of Gothenburg

2012
Outline

Background

Swedish language history

Lexica

Corpora

Lexical link-up
Why historical resources?

Questions:

► How did *do* support develop in English?
► How was the present participle used in Old Swedish?
► 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.’

English *do*-support

Kroch (1989), data from Ellegård
Present participle in Old Swedish

How was the present participle used in Old Swedish? Can we distinguish influences from Latin or German? [Ahlberg, 1942]:

‘Native’ use:
   a) kom gangande ‘come walking’, skinande huita ‘shining white’
   b) han gick söriande wtu kammaren ‘he left the room sad’

Innovation, poss due to Latin, in translated texts:
   c) P yplypte han sighande...
      P elevavit eum dicens...
      ‘Peter lifted him up and said...’

Not everything goes!
   d) the hörande wetta thorkadhos j sinom hiærom audientes autem haec, dissecabantur cordibus suis
      ‘When hearing this, they were cut to the heart (hurt/offended)...’
The change of *experimental*

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

- *experiment*: religious $\rightarrow$ 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

### Background

- Swedish language history
- Lexica
- Corpora
- Lexical link-up

### Swedish Lexica

<table>
<thead>
<tr>
<th>Swedish</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>uppmöblera</td>
<td>sofa</td>
</tr>
<tr>
<td>kanapé</td>
<td>seat</td>
</tr>
<tr>
<td>bärstol</td>
<td>armchair</td>
</tr>
<tr>
<td>valkbord</td>
<td>table</td>
</tr>
<tr>
<td>tryckbord</td>
<td>print</td>
</tr>
<tr>
<td>tingsbord</td>
<td>table</td>
</tr>
<tr>
<td>tebord</td>
<td>tea table</td>
</tr>
<tr>
<td>svärtbord</td>
<td>blackboard</td>
</tr>
<tr>
<td>strykbord</td>
<td>ironing board</td>
</tr>
<tr>
<td>stigbord</td>
<td>footstool</td>
</tr>
<tr>
<td>skådebord</td>
<td>cellar</td>
</tr>
<tr>
<td>skärbord</td>
<td>desk</td>
</tr>
<tr>
<td>skänkbord</td>
<td>armchair</td>
</tr>
<tr>
<td>ljusbord</td>
<td>lamp table</td>
</tr>
<tr>
<td>kredensbord</td>
<td>credenza</td>
</tr>
<tr>
<td>kammarbord</td>
<td>dressing table</td>
</tr>
<tr>
<td>hästskobord</td>
<td>horsehair</td>
</tr>
<tr>
<td>herrskapsbord</td>
<td>table</td>
</tr>
<tr>
<td>färög bord</td>
<td>feudal lord</td>
</tr>
<tr>
<td>friserbord</td>
<td>hairdresser's table</td>
</tr>
<tr>
<td>fortunabord</td>
<td>fortune table</td>
</tr>
<tr>
<td>formbord</td>
<td>table</td>
</tr>
<tr>
<td>dambord</td>
<td>drum table</td>
</tr>
<tr>
<td>brödstbord</td>
<td>breadboard</td>
</tr>
<tr>
<td>brädspelsbord</td>
<td>games table</td>
</tr>
<tr>
<td>bord</td>
<td>table</td>
</tr>
<tr>
<td>bakbord</td>
<td>backboard</td>
</tr>
<tr>
<td>altabord</td>
<td>altar board</td>
</tr>
<tr>
<td>tidningsbyrå</td>
<td>newspaper office</td>
</tr>
<tr>
<td>ettoman</td>
<td>armchair</td>
</tr>
<tr>
<td>kommissionsbyrå</td>
<td>commission office</td>
</tr>
<tr>
<td>klädesbyrå</td>
<td>clothes office</td>
</tr>
<tr>
<td>divan</td>
<td>sofa</td>
</tr>
<tr>
<td>byrå</td>
<td>bureau</td>
</tr>
<tr>
<td>bord</td>
<td>table</td>
</tr>
<tr>
<td>affärsbyrå</td>
<td>business office</td>
</tr>
<tr>
<td>adressbyrå</td>
<td>address office</td>
</tr>
<tr>
<td>vändstol</td>
<td>revolving chair</td>
</tr>
<tr>
<td>väggbänk</td>
<td>backboard</td>
</tr>
<tr>
<td>vridstol</td>
<td>revolving chair</td>
</tr>
<tr>
<td>verkbänk</td>
<td>workbench</td>
</tr>
<tr>
<td>vaskbänk</td>
<td>washstand</td>
</tr>
<tr>
<td>varpstol</td>
<td>kip stool</td>
</tr>
<tr>
<td>varmbänk</td>
<td>sunbath</td>
</tr>
<tr>
<td>valstol</td>
<td>value table</td>
</tr>
<tr>
<td>vaktbänk</td>
<td>guard bunck</td>
</tr>
<tr>
<td>understol</td>
<td>under stool</td>
</tr>
<tr>
<td>tvättstol</td>
<td>washing stand</td>
</tr>
<tr>
<td>tvärbänk</td>
<td>parallel board</td>
</tr>
<tr>
<td>torfbänk</td>
<td>peat bank</td>
</tr>
<tr>
<td>sågbänk</td>
<td>saw bench</td>
</tr>
<tr>
<td>svartbänk</td>
<td>blackboard</td>
</tr>
<tr>
<td>strumpstol</td>
<td>stocking stool</td>
</tr>
<tr>
<td>stol</td>
<td>table</td>
</tr>
<tr>
<td>squbänk</td>
<td>squab</td>
</tr>
<tr>
<td>spegelbänk</td>
<td>mirror</td>
</tr>
<tr>
<td>sofjstol</td>
<td>sofa</td>
</tr>
<tr>
<td>slibbänk</td>
<td>slipper stand</td>
</tr>
<tr>
<td>slagtbänk</td>
<td>slagboard</td>
</tr>
<tr>
<td>slatarbänk</td>
<td>slatted backboard</td>
</tr>
<tr>
<td>skårbänk</td>
<td>cupboard</td>
</tr>
<tr>
<td>skottstol</td>
<td>skittler stool</td>
</tr>
<tr>
<td>skjutstol</td>
<td>shawl stool</td>
</tr>
<tr>
<td>röstand</td>
<td>rolling stool</td>
</tr>
<tr>
<td>ryggstol</td>
<td>backrest</td>
</tr>
<tr>
<td>nulibänk</td>
<td>null bank</td>
</tr>
<tr>
<td>kardbänk</td>
<td>card table</td>
</tr>
<tr>
<td>kalkbänk</td>
<td>limeboard</td>
</tr>
<tr>
<td>häftstol</td>
<td>writing table</td>
</tr>
<tr>
<td>hvilostol</td>
<td>rest stool</td>
</tr>
<tr>
<td>huggbänk</td>
<td>hammer bench</td>
</tr>
<tr>
<td>hjustol</td>
<td>horse table</td>
</tr>
<tr>
<td>handstol</td>
<td>hand table</td>
</tr>
<tr>
<td>halmstol</td>
<td>hay table</td>
</tr>
<tr>
<td>gubbstol</td>
<td>gull table</td>
</tr>
<tr>
<td>grässtol</td>
<td>grass stool</td>
</tr>
<tr>
<td>fogbänk</td>
<td>fog board</td>
</tr>
<tr>
<td>fallbänk</td>
<td>fallboard</td>
</tr>
<tr>
<td>erkebiskopsstol</td>
<td>bishop's throne</td>
</tr>
<tr>
<td>duftstol</td>
<td>scent table</td>
</tr>
<tr>
<td>dragstol</td>
<td>drag stool</td>
</tr>
<tr>
<td>bönstol</td>
<td>beetle stool</td>
</tr>
<tr>
<td>bänk</td>
<td>bench</td>
</tr>
<tr>
<td>brudbänk</td>
<td>bride's stool</td>
</tr>
<tr>
<td>borstol</td>
<td>borage stool</td>
</tr>
<tr>
<td>bordbänk</td>
<td>board bench</td>
</tr>
<tr>
<td>bokstol</td>
<td>book table</td>
</tr>
<tr>
<td>biskopsstol</td>
<td>bishop's throne</td>
</tr>
<tr>
<td>bitbänk</td>
<td>bit table</td>
</tr>
<tr>
<td>bandstol</td>
<td>band table</td>
</tr>
<tr>
<td>armstol</td>
<td>arm table</td>
</tr>
<tr>
<td>vaksäng</td>
<td>wax bed</td>
</tr>
<tr>
<td>trädgårdssäng</td>
<td>garden bed</td>
</tr>
<tr>
<td>säng</td>
<td>bed</td>
</tr>
<tr>
<td>syskonsäng</td>
<td>sibling bed</td>
</tr>
<tr>
<td>sparringsäng</td>
<td>lovers' bed</td>
</tr>
<tr>
<td>skogsäng</td>
<td>woods bed</td>
</tr>
<tr>
<td>sjuksäng</td>
<td>sick bed</td>
</tr>
<tr>
<td>plantsäng</td>
<td>plant bed</td>
</tr>
<tr>
<td>paulunsäng</td>
<td>Paul's bed</td>
</tr>
<tr>
<td>neglssäng</td>
<td>negl bed</td>
</tr>
<tr>
<td>melonsäng</td>
<td>melon bed</td>
</tr>
<tr>
<td>lösäng</td>
<td>loa bed</td>
</tr>
<tr>
<td>krydsäng</td>
<td>cross bed</td>
</tr>
<tr>
<td>korgsäng</td>
<td>basket bed</td>
</tr>
<tr>
<td>gurksäng</td>
<td>gorilla bed</td>
</tr>
<tr>
<td>faitsäng</td>
<td>feet bed</td>
</tr>
<tr>
<td>dödssäng</td>
<td>death bed</td>
</tr>
<tr>
<td>bröllopssäng</td>
<td>wedding bed</td>
</tr>
<tr>
<td>biomsäng</td>
<td>bone bed</td>
</tr>
<tr>
<td>biomstersäng</td>
<td>bone bed</td>
</tr>
<tr>
<td>bergsäng</td>
<td>mountain bed</td>
</tr>
<tr>
<td>utdragssoffa</td>
<td>pull-out sofa</td>
</tr>
<tr>
<td>soffa</td>
<td>sofa</td>
</tr>
<tr>
<td>hvilosoffa</td>
<td>leisure sofa</td>
</tr>
</tbody>
</table>

Figure 3: Words semantically related to *soffa* ‘sofa’

[Borin et al., 2011]
Historical linguistics

- Historical texts form a closed corpus of data
- Traditional differences corpus vs historical linguistics:
  - Corpus linguistics: strive for representativeness
  - Historical linguistics: selective approach to data, focus on particular phenomena
- Historical linguistics is now turning towards corpus linguistics
- Creating larger historical corpora has thus become interesting
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
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.
Swedish language history

Indo-European > Germanic > North G. > Swedish

Proto-Germanic
Proto-Norse
Rune-Swedish
Early Old Swedish
Late Old Swedish
Early Modern Swedish
Late Modern Swedish
Contemporary Swedish
Swedish language history

Indo-European > Germanic > North G. > Swedish

**Proto-Germanic**
- Proto-Norse
- Rune-Swedish
- Early Old Swedish
- Late Old Swedish
- Early Modern Swedish
- Late Modern Swedish
- Contemporary Swedish

- ca 750-250 BC
- Germanic sound shift (Grimm’s law): pater → father
- accent → first syllable
Swedish language history

Indo-European > Germanic > North G. > Swedish

- Proto-Germanic
  - Proto-Norse
  - Rune-Swedish
- Early Old Swedish
- Late Old Swedish
- Early Modern Swedish
- Late Modern Swedish
- Contemporary Swedish

- ca 150-800
- Syncope (loss of unstressed vowel) due to accent shift: gastiR → gästR
- definite article after noun: the house vs huset
- West vs East Norse (Swedish)
Swedish language history

Indo-European > Germanic > North G. > Swedish

Proto-Germanic
Proto-Norse
**Rune-Swedish**
Early Old Swedish
Late Old Swedish
Early Modern Swedish
Late Modern Swedish
Contemporary Swedish

- ca 800-1225
- diphtongs → monophtongs
Swedish language history

Indo-European > Germanic > North G. > Swedish

Proto-Germanic
Proto-Norse
Rune-Swedish
Early Old Swedish
Late Old Swedish
Early Modern Swedish
Late Modern Swedish
Contemporary Swedish

- ca 1225-1375
- fairly stable language
- Äldre västgötalagen 1225
Swedish language history

Indo-European > Germanic > North G. > Swedish

Proto-Germanic
Proto-Norse
Rune-Swedish
**Late Old Swedish**
Early Old Swedish
Early Modern Swedish
Late Modern Swedish
Contemporary Swedish

- ca 1375-1526
- Language change
- Influence from Low German
- Ends with book printing (Gustav Vasa’s Bible 1526/1541)
Swedish language history

<table>
<thead>
<tr>
<th>Indo-European</th>
<th>Germanic</th>
<th>North G.</th>
<th>Swedish</th>
</tr>
</thead>
</table>

- **Proto-Germanic**
- **Proto-Norse**
- **Rune-Swedish**
- **Early Old Swedish**
- **Late Old Swedish**
- **Early Modern Swedish**
- **Late Modern Swedish**
- **Contemporary Swedish**

- ca 1526-1732
- Written standard emerges
- Formal church language vs everyday language
- Ends with first issue of Olof von Dalin’s 'Svenska Argus'

Background

Swedish language history

Lexica

Corpora

Lexical link-up
Swedish language history

Indo-European > Germanic > North G. > Swedish

Proto-Germanic
Proto-Norse
Rune-Swedish
Early Old Swedish
Late Old Swedish
Early Modern Swedish
Late Modern Swedish
Contemporary Swedish

- ca 1732-1906
- Standard orthography and stable grammar
- Ends with the latest Swedish language reform
An historical infrastructure for Swedish

- **Historical lexica**
  - Schlyter (13th–16th c. law)
  - Söderwall & Söderwall supplement (13-16th c.)
  - Swedberg (18th c.)
  - Dalin (19th c.)

- **Historical corpora**
  - More and more texts are being digitized
  - Fornsvenska textbanken, Project Runeberg, Riksantikvarieämbetet
  - HaCossa [Höder, 2011]

- **Historical morphology**
  - [Borin and Forsberg, 2008]
Lexical resources

How does this become computational?

- OCR/double keying
- XML
- Entry identifiers
- Sense identifiers
Lexical resources

http://spraakbanken.gu.se/karp/ #lang=en
Corpora development

- Sentence splitting, tokenization
- Part-of-speech tagging
- (Lemmas, morphological annotation)
- (Parsing)
- (Semantic annotation)
- (Named entity recognition)
- (Alignment)
- (….)
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.)
Explore the Old Swedish text!

How can we convert this into a corpus?
What could be problematic?
What could be easy?
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
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 mønstrit 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 mønstrit 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 mønstersins 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
Pre-processing and tagging

Tokenization and sentence splitting

- Pre-processing required by most annotation tools
- Standard tools may not work

Part-of-speech and morphology

- Taggers often use lexical info and statistics
- Problems: spelling variation, different vocabulary and morphology
- Common approaches:
  - Normalize to modern spelling and use available tools
e.g., [Pettersson et al., 2012], [Hinrichs and Zastrow, 2012]
  - Manual annotation and training new model
e.g., [Dipper, 2011]
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.

Recursive definition of Levenshtein distance (‘standard’ algorithm, [Wagner and Fisher, 1974], many others):

\[
\begin{align*}
    d_L(\epsilon, O) &= |O| \\
    d_L(I, \epsilon) &= |I| \\
    d_L(Ix, Oy) &= \begin{cases} 
        d_L(I, O) & \text{when } x = y \\
        1 + \min \left( d_L(Ix, O), d_L(I, Oy), d_L(I, O) \right) & \text{when } x \neq y 
    \end{cases}
\end{align*}
\]

Variants using fewer operation types (Hamming, Longest common substring); more operation types (’transpose’: Damerau); \(n\rightarrow m\) rules; weighted rules (generalized edit distance).
Levenshtein Distance

Also have a look at the explanation and implementation at http://odur.let.rug.nl/kleiweg/lev/
Background

Swedish language history

Lexica

Corpora

Lexical link-up

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]:

<table>
<thead>
<tr>
<th>Rule</th>
<th>Wght</th>
<th>Schlyter example</th>
</tr>
</thead>
<tbody>
<tr>
<td>( u \rightarrow o )</td>
<td>0.20</td>
<td>\textit{arvöpi} ( ærvöpi ) ‘work’</td>
</tr>
<tr>
<td>( æ \rightarrow e )</td>
<td>0.27</td>
<td>\textit{æer} ( er ) ‘scar’</td>
</tr>
<tr>
<td>( pt \rightarrow ft )</td>
<td>0.31</td>
<td>\textit{apter} ( after ) ( æftær ) ‘after’</td>
</tr>
<tr>
<td>( g# \rightarrow gg# )</td>
<td>0.42</td>
<td>\textit{væg} ( vægg ) ( vegg ) ‘wall’</td>
</tr>
<tr>
<td>( ðer \rightarrow n )</td>
<td>0.43</td>
<td>\textit{maðer} ( man ) ‘man’</td>
</tr>
<tr>
<td>( au \rightarrow ö )</td>
<td>0.44</td>
<td>\textit{barnlös} ( barnalös ) ( barnalaus ) ‘childless’</td>
</tr>
<tr>
<td>( th \rightarrow b )</td>
<td>0.44</td>
<td>\textit{opolskipti} ( oðalskipte ) ‘(type of) land redivision’</td>
</tr>
<tr>
<td>( mp \rightarrow m )</td>
<td>0.45</td>
<td>\textit{hamn} ( hampn ) ‘harbour’</td>
</tr>
<tr>
<td>( li \rightarrow eli )</td>
<td>0.45</td>
<td>\textit{lastelika} ( lastlika ) ‘blameworthy, shameful’</td>
</tr>
<tr>
<td>( ghi \rightarrow i )</td>
<td>0.62</td>
<td>\textit{aplöja} ( opplöghia ) ‘plowing into the neighbour’s field’</td>
</tr>
</tbody>
</table>
Normalization strategies

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

\[
\begin{align*}
    h & \rightarrow \epsilon / \# _ _ v \\
    h & \rightarrow \epsilon / g _ \\
    k & \rightarrow g / V _ V \\
    i & \rightarrow j / [t k s] _ [\ddot{a} y u \ddot{o} e] \\
    u & \rightarrow o / _ _ \# \\
    qu & \rightarrow kv \\
    \text{etc.}
\end{align*}
\]

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

\[
\begin{align*}
    [w v u o h v h u] & \rightarrow u \\
    [i j ii ij] & \rightarrow i \\
    [\ddot{a} \dddot{a} e] & \rightarrow e \\
    [h] & \rightarrow \epsilon \\
    \text{etc.}
\end{align*}
\]
Normalization strategies

Soundex ([Knuth, 1973], US patent from 1918 by O’Dell & Russel!) and many, many variants, creates a hash based upon (much simplified) pronunciation:

- Convert characters according to table, remove repeats, exchange first digit for first character, remove zeroes, adjust length.


Carl Bloms Boktryckeri, Lund.


In *Proceedings of the LREC 2008 Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2008)*, pages 9–16, Marrakech, Marocco. ELRA.

References II


References III

Phrases and clauses tagging manual for syntactic analyses of old nordic texts encoded as menotic xml documents (pacman). version 2.0.

The Art of Computer Programming: Volume 3, Sorting and Searching.
Addison-Wesley.
Knuth's version of the Soundex algorithm is described in Ch 6.

Function and grammar in the history of English: periphrastic do.

Parsing the past - identification of verb constructions in historical text.
References IV


