Towards A Field of Comparative Quantitative Information Structure

Joel C. Wallenberg
University of Iceland
joel.wallenberg@gmail.com

SHES 9
April 17, 2011
Universiteit Leiden
Outline

1. Why “Comparative Quantitative”? 
   - Basic Questions
   - Parallel Parsed Corpora

2. Intro to Information Structure

3. Passive
   - Frequency of Passives in Icelandic, German, and English

4. NEG Topicalization
   - Comparing German and Icelandic
   - Comparing Icelandic and English
   - A Change Over Time in Icelandic?

5. Conclusions
Basic Questions

- A speaker uses syntax and prosody in order to organize information for a hearer (Information Structure).
- How does Information Structure exercise different syntactic options in order to do this?
- How does IS interact with syntax differently in different languages with different syntactic constraints?
- How can we generate and test such hypotheses rigorously?
Basic Answers

- We can rely on constructed data, intuitions, and experimentation.
- We can use production data.
  - Collected naturally occurring examples are difficult to interpret in terms of information structure, because of a need to control context.
  - Collecting naturally occurring examples in order to compare different languages is even more difficult, because of the need to control context (and other things) across languages.
  - It is difficult to find what you want, for any specific phenomenon under study.
- Corpus data: it *might* easier to find what you want, but the other problems apply.
Parallel Parsed Corpora

- The Icelandic Parsed Historical Corpus (IcePaHC) (Wallenberg et al., 2011)
  - currently 444,369 words (as of Thursday)
- The Penn Parsed Corpus of Early Modern English (Kroch et al., 2004)
  - ~1.8 million words
- The Parsed Corpus of Early New High German (Light, 2011)
  - currently ~70,000 words.
- Each of these has a parsed sample of the New Testament, which includes the Gospel of John (~20,000 words)
  - Oddur Gottskálksson, date: 1540
  - William Tyndale, date: 1525/1534
  - Martin Luther *Septembertestament*, date: 1522
Parallel Parsed Corpus of the New Testament

- Translations of the same text, but not slavish ones.
  - Protestant Bible translations were meant to be read by normal people.
  - The timing of the translations means that the translation influence is mostly from Luther.

- You can control for context, because the texts are conveying the same information in every verse.

- You can search for specific constructions in any of the languages, in order to compare with the others.
  - Especially constructions which have a particular function, and are known to be ungrammatical in one or more of the languages.

- Frequency information.
2. Intro to Information Structure

- The speaker directs the hearer to pay attention to certain information, and store it with reference to other information.

- The **topic** is “an address pointer in the sense that it directs the hearer to a given address..in the hearer’s knowledge-store, under which the information [JCW: i.e. **focus**] carried by the sentence in entered”(Vallduví, 1992, 52)

- The **topic** is an “open proposition” which the speaker presupposes to be the true in the world, and it contains a variable which is instantiated by the **focus** constituent of the sentence. (Prince, 1999, 6)
Topic and Focus

- **VP/νP (Wide) Focus:**
  
  1. **Question:** What did John do?  
  **Answer:** John interrupted Mary at the SHES workshop.

  - **Topic:** John VPed.
  - **Focus:** interrupted Mary at the SHES workshop.
Object (Narrow) DP Focus:

(2) Question: Who did John interrupt at the SHES workshop?
Answer: (John interrupted) Mary (at the SHES workshop).

- Topic: John interrupted X at SHES.
- Focus: Mary.

Languages mark Topic and Focus in various morphosyntactic and/or prosodic ways, e.g. in Germanic with the nuclear accent (final pitch accent) on Focus (cf. Ladd, 1996, *inter alia*).
3. Passive: Passives in John

- Joint work with Caitlin Light (U of PA).

<table>
<thead>
<tr>
<th></th>
<th>Passive</th>
<th>Active</th>
<th>Freq. Passive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tyndale</td>
<td>146</td>
<td>914</td>
<td>0.138</td>
</tr>
<tr>
<td>Luther</td>
<td>108</td>
<td>1071</td>
<td>0.092</td>
</tr>
<tr>
<td>Oddur</td>
<td>86</td>
<td>1100</td>
<td>0.073</td>
</tr>
</tbody>
</table>

EME vs. ENHG: Chi-square = 11.3571 on 1df, \( p = 0.0008 \)
EME vs. Icelandic: Chi-square = 25.008 on 1df, \( p \approx 0 \)
ENHG vs. Icelandic: Chi-square = 2.6136 on 1df, \( p = 0.1059 \)
The Effect is Real

- Results reproduced over full corpora of Early Modern English (PPCEME) and Icelandic (IcePaHC v0.4).

<table>
<thead>
<tr>
<th></th>
<th>Passive</th>
<th>Active</th>
<th>Freq. Passive</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPCEME</td>
<td>17273</td>
<td>40096</td>
<td>0.301</td>
</tr>
<tr>
<td>IcePaHC</td>
<td>2403</td>
<td>14467</td>
<td>0.142</td>
</tr>
</tbody>
</table>

Chi-square = 1683.547 on 1df, \( p \approx 0 \)
Why is there a difference?

- In Luther, there are 54 non-passive tokens corresponding to passives in Tyndale.
- 7 of these are not full translations, but 41 out of the remaining 48 (85.4%) have the same structural subject as Tyndale’s passives.
- Luther uses various non-passive constructions to make the Theme the structural subject, most frequently copular constructions with adjectival predicates (11/48).
In Oddur Gottskálksson, there are 77 non-passive tokens corresponding to passives in Tyndale.

Of the 77, 50 (64.9%) correspond to –st middle verbs in Icelandic.

11 (14.3%) correspond to copular constructions with adjectival predicates.

Only 9 (11.7%) correspond to actives. (And then there are 7 examples of other types of constructions.)

**Note:** 72 out of 77 (93.5%) non-corresponding clauses have the same referent as the structural subject as Tyndale does, including 49/50 of the middles and 8/9 of the actives.
Why is there a difference?

- **Observation:** All three authors in all three languages encode the same entity as the Subject.
  - However, the syntactic means of doing so (or derivational history) does not seem to matter.

(3) **John 3:23**

a. and they came and were baptised (Tyndale)

b. vnd sie kamen dahynn vnd ließen sich teuffen and they came there and let REFL baptize “And they came there and had themselves baptized” (Luther)

C. Þeir komu þangað og skírðust They came thence and baptized-MID. (Oddur)
Why is there a difference?

(4) **John 16:20**

a. Ye shall sorowe: but youre sorowe shalbe tourned to ioye (Tyndale)

b. doch ewr traurickeyt soll zur freud werden but your sorrow shall to joy become. (Luther)

c. ér verðið og hryggvir, en yðar hryggð skal you become also sad, but your sorrow shall snúast í fögnuð. turn-MID into joy. (Oddur)
Observation: All three authors in all three languages encode the same entity as the Subject.

Inference: Some syntactic notion of Subject is used to encode the same information structural element in all three languages.

- Note: since the V2 languages do not use fronted objects for this purpose, this is likely not reducible to string-wise position.
- Note: these languages do not necessarily have the same subject position(s), and certainly not the same derived subject position.

Hypothesis: The relevant IS role is Topic (see Vallduví 1992, and work in Centering Theory), and the relevant syntactic notion is the highest available A-spec.
One addition question: Passives in Historical English

- Three prose translations of the Rule of St. Benet:
  Old English (11th c.), Northern Middle English (1425),
  Southern/Kentish Middle English (1490).
  - PennParsed Corpus of Middle English (Kroch and Taylor, 2000)

<table>
<thead>
<tr>
<th></th>
<th>Passive</th>
<th>Active</th>
<th>Freq. Passive</th>
</tr>
</thead>
<tbody>
<tr>
<td>OE Rule</td>
<td>348</td>
<td>871</td>
<td>0.286</td>
</tr>
<tr>
<td>North ME Prose</td>
<td>218</td>
<td>514</td>
<td>0.298</td>
</tr>
<tr>
<td>William Caxton’s</td>
<td>90</td>
<td>97</td>
<td>0.481</td>
</tr>
</tbody>
</table>

OE vs. North ME: Chi-square $= 0.281$ on 1df, $p = 0.59$
OE vs. North ME vs. South ME: Chi-square $= 29.6$ on 2df, $p \approx 0$
4. NEG Topicalization

- Looked at this configuration in Luther:
  Auxiliary > Negation/Adverbial > DP Object > Lexical-verb
- Allows default nuclear accent to fall on the DP object, which can be DP focus or VP focus.
- No scrambling of the object leftward, which would move the object away from the default nuclear accent (Lenerz, 1977, and subsequent work).

(5) Jhr habt nie widder seyne stym gehoret noch
You have not neither his voice heard nor
seyne gestalt gesehen
his form seen
“You have neither heard his voice nor seen his form” (John V, 37)
Scrambling in German

- Auxiliary > Negation/Adverbial > DP Object > Lexical-verb
- There were only 7 matrix clauses with this exact configuration I looked for in John.
- However, 5 out of 7 of these examples (71%) in German correspond to topicalization of either negation or the adverbial in Oddur Gottskálksson’s version of John.

(6) Eigi hafi þér nokkurn tíma heyrt hans rödd né
    Not have you some time heard his voice nor
    séð hans ásýnd
    seen his form
    “You have at no time heard his voice nor seen his form” (John V, 37)
Because of this correspondence, I did a study of NEG topicalization in the Oddur.

This is ungrammatical in both German and English, so they must use some different strategy.

Observation: there are 2 types of NEG topicalization in Oddur’s language.

1. NEG itself is focused (and probably accented).
2. NEG is not focused; it has been moved out of the prosodic domain of the focus.
**NEG Topicalization**

- 20 examples in matrix, non-copular clauses.

<table>
<thead>
<tr>
<th>Focus on:</th>
<th>Number of Matrix Clauses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negation</td>
<td>2</td>
</tr>
<tr>
<td>Object or VP-internal</td>
<td>7</td>
</tr>
<tr>
<td>VP</td>
<td>6</td>
</tr>
<tr>
<td>Verb</td>
<td>3</td>
</tr>
<tr>
<td>Subject</td>
<td>2</td>
</tr>
<tr>
<td>Total:</td>
<td>20</td>
</tr>
</tbody>
</table>
Two Types of NEG Topicalization

(7) **Focus on NEG:**
Þá svöruðu Júðar og sögðu honum: Segju vær eigi vel að þú ert samverskur og hefir djöfulinn?

Eigi hefi eg djöful
Not have I devil

“Do we not say right that you are a Samaritan and have the Devil? ...I don’t have any devil.” (John VIII, 49)

(8) **Focus on VP:**
þetta var og gjört (P:230) svo að ritningin uppfylltist:

Eigi skulu þér brjóta nokkuð hans bein.
Not shall you break some his bone

“This was also done so that the scripture was fulfilled: you shall not break any of his bones.” (John XIX, 36)
Luther’s treatment of the same 20 sentences (all NEG top. in Oddur):

<table>
<thead>
<tr>
<th>Focus on</th>
<th>NEG &gt; Foc</th>
<th>Neg. Object</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEG</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Obj/VP-internal</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>VP</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Verb</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Subj</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Total:</td>
<td>9</td>
<td>6</td>
<td>5</td>
<td>20</td>
</tr>
</tbody>
</table>

In 15/20 (75%) sentences, Luther places negation to the left of Focus or removes negation.
German (Luther) vs. Early Modern English (Tyndale)

- Tyndale’s treatment of the same 20 sentences (all NEG top. in Oddur).
- Since English is VO, Neg > Foc is expected for Object Focus, and for VP focus with an auxiliary.

<table>
<thead>
<tr>
<th>Focus on:</th>
<th>NEG &gt; Foc</th>
<th>Neg. Object</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEG</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Obj/VP-internal</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>VP</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Verb</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>11</td>
<td>4</td>
<td>3</td>
<td>18</td>
</tr>
</tbody>
</table>

- As in Luther, Tyndale places negation to the left of the focus or removes it in 15 sentences.
**Neither-Topicalization**

- In the remaining 2 Subject Focus sentences, Tyndale translates the negation with topicalized *Neither*:

(9) Woman where are those thyne accusars? Hath no man condempned the? She sayde: No man Lorde. And Iesus sayde: Nether do I condempne the.

   (John VIII, 11)

(10) Master who dyd synne: this man or his father and mother that he was borne blynde? Iesus answered: Nether hath this man synned nor yet his father and mother.

   (John IX, 3)

- Tyndale uses this construction only 3 other times, but they all correspond to NEG top. in Oddur (in subordinate clauses).
Modern Icelandic

- It is unclear if speakers of modern Icelandic accept NEG topicalization that does not focus negation.
- So, 11 is possible in the “Hver er maðurinn?” (20 questions) game. (Thank you to Einar Freyr Sigurðsson for these examples)

(11) Q:
Dó hann í fyrra?
Died he in yesteryear?
“Did he die last year”?
A:
Ekki dó hann í fyrra.
NEG died he in yesterday.
“He did not die last year.”
Modern Icelandic

- In the game example, there is clear focus on NEG.
- In fact, the answer could have just been *Nei* ("No"), so NEG is really the only information provided.
- Focus accompanied by clear accent on “ekki”.
- But examples with object or VP focus seem to be less acceptable.
Modern Icelandic: Wide (VP) Focus

(12) A: Við fórum í Kringluna í gær.
    We went to Kringlan-the yesterday

B: Hvernig var?
    How was

A:

a. Við fundum ekki neitt og fórum heim.
    We found NEG anything and went home

b. ?? Ekki fundum við neitt ...
    NEG found we anything ...

Modern Icelandic: Object (DP) Focus

(13)  A: Við fórum í Kringluna í gær.
We went to Kringlan-the yesterday

B: Hvað keyptirðu?
What bought-you?

A:

a. Óg keypti ekki neitt.
I bought NEG anything.

b. ?? Ekki keypti ég neitt.
NEG bought I anything.
A Change Over Time in Icelandic?

Topicalization of Negation Over Time

- Frequency of Topicalization
- Year:
  - 1100
  - 1200
  - 1300
  - 1400
  - 1500
  - 1600
  - 1700
  - 1800

Graph showing the frequency of topicalization of negation over time from 1100 to 1800.
A Change in Icelandic?

Topicalization of Negation Over Time

- Blue dots: Clauses With Auxiliaries
- Red dots: Clauses with Finite Lexical Verbs

Year
- 1100
- 1200
- 1300
- 1400
- 1500
- 1600
- 1700
- 1800

Frequency of Topicalization
- 0.0
- 0.2
- 0.4
- 0.6
- 0.8
- 1.0
Conclusions

- One A-movement technique for encoding IS, and one A’-movement.
- In the use of the passive, German and Icelandic pattern together to some extent (vs. English), but each language has different morphosyntactic resources.
- However, while English differs in its use of passive, the languages agree on the use of the structural subject.
- Why do OE and Northern ME pattern together? Could V2 have to do with this?
Conclusions

- NEG topicalization in (Oddur’s) Icelandic has multiple functions:
  - To Focus NEG, or to move NEG away from the Focus.
  - Since the latter subsumes a number of Focus configurations, I suggest its motivation is prosodic.
  - The change in Icelandic might make sense according to a prosodic explanation.

- More research is necessary.

- None of this research is possible without parsed corpora, and in particular, parallel parsed corpora.
References I


