The use of parsed corpora in information structural research

*LSA Summer Institute 2013: Workshop on Diachronic Syntax*

Caitlin Light

University of York

June 29, 2013
Introduction

- We have already seen some discussion of the use of quantitative data from parsed corpora for syntactic research.
  - Perhaps a more challenging question is whether corpus data can (and should) be used in investigating information structure.
  - This has been an issue of some controversy in recent work.

- Today I will consider some of the issues and advantages related to the use of corpus data in information structural inquiry.
  - Corpus data can be key to pushing our understanding of information structure forward, but only if used carefully.
  - A case study on passivization in the history of English demonstrates a possible methodology for this type of research.
Outline of the talk

1. **Information structural research and corpora**
   - Difficulties and issues
   - The importance of corpus data
   - First steps

2. **Case study**
   - Passives in the history of English
   - A link between passives and V2?

3. **Investigating the question**
   - Comparing English and its closest relatives
   - Comparing stages of English
   - Parallel parsed corpora
   - The Rule of St. Benedict verse comparison
   - The New Testament verse comparison

4. **Conclusion**
   - Structure of the investigation
Basic questions

- A speaker uses syntax and prosody in order to organize information for a hearer (Information Structure).
- How does IS manipulate syntax in order to do this?
- How does IS interact with syntax differently in different languages with different syntactic constraints? Furthermore, what remains the same?
- How can we generate and test such hypotheses rigorously?
Some possible 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?
Difficulties and issues

- The utilization of corpus data for information structural research is not necessarily straightforward.
  - Most existing parsed corpora are not annotated for information structural information.

- In fact, attempts to annotate for information structural categories have met with a variety of challenges (Bech, 2013; Cook, 2013).
  - Information structural annotations are found to be inconsistent.
  - Studies suggest that we require a deeper theoretical understanding to properly implement them.

- Information structure is a relatively young subfield, and many of the problems may come from the attempt to apply pre-theoretical assumptions to quantitative data.
However, corpora offer massive numbers of naturally occurring examples (in certain registers).

- This has some of the disadvantages of production data: in particular, we cannot control for context.
- But **parallel parsed corpora** may help solve the problem of cross-linguistic study.

Furthermore, because parsed corpora are pre-existing resources, they can provide a data set not biased by the researcher’s expectations.
I argue that corpus data can help **shed light** on our existing questions about information structure.

- We must find methods of investigating information structure in corpora **without** relying on pre-theoretical notions.
- Our theories must then be built around the evidence.

The following case study is intended as an illustration of how such an investigation could be structured.

- We begin by comparing the **syntactic constructions**, independent of any assumptions about their information structure.
The overall rate of passivization in English has risen significantly since the Old English period.

- What’s more, we see the appearance of new passive-like constructions, like the so-called prepositional passive.

Los (2002, 2009); Seoane (2006) suggest that the rise in English passive has an information structural cause.

- As English lost certain word order options, other word orders were commandeered to accomplish the same information structural goal.

This case study will consider their claims in the light of new quantitative data (Light and Wallenberg, 2011).
Passivization and V2 topicalization as IS equivalents?

Recent work on the syntax/information structure interface introduces the proposal that unaccented V2 topicalization and passivization have an information structurally equivalent effect on the topicalized or promoted object, particularly in the history of English.

(1) Matthew 13:27–28
   a. Herre, hastu nit guten samen auff deynen acker
      Lord have-you not good seeds on your acre
geseet? wo her hatt er denn das vnkraut? vnd
sowed where from has he then the weeds and
er sprach, das hat eyn feyndt than
he spoke this has an enemy done
   b. This was done by an Enemy.
      (our constructed example)
V2-like word orders in Old English

- As we saw this morning, Old English was not a V2 language in the way German is.
  - However, it did allow V2-like word orders with topicalized objects, which are no longer generally permitted in Modern English.

- Speyer: (unaccented) personal pronouns can topicalize in Old English, but rate of pronoun topicalization rapidly declines in Middle English.

(2) **Pone** asende se Sunu
  this sent the son
  ‘The son sent this one.’
  (coaelhom,+AHom_9:113.1350)

(3) **hit** Englisse men swy3e amyrdon
  and it English men fiercely prevented
  ‘and the Englishmen prevented it fiercely.’
  (cochronE,ChronE_[Plummer]:1073.2.2681)
Passivization and V2 topicalization as IS equivalents?

As Historical English loses the ability to generate V2 word orders, and the language becomes more rigidly SVO, passivization becomes the preferred construction to promote a non-subject argument to a high, unaccented position (cf. Los, 2009; Seoane, 2006).

- Argument: both unaccented topicalization and promotion of an argument in the passive result in marking the DP as informationally topical/thematic.
- Thus, the rise in passivization can be seen as a strategy to compensate for the loss of V2 word orders.

This argument has intuitive appeal, and the information structural/pragmatic claims have much support in the general literature on these constructions.
A quantitative study in Seoane (2006)

- Seoane (2006) presents a corpus-based study of this phenomenon in late Middle and Early Modern English.
- Passives with by-phrases are considered for their informational content.
  - Both the promoted subject and the demoted agent are coded for definiteness, givenness, human/animacy, and other properties thought to be characteristic of topics.
  - These are then compared to determine whether passivization does indeed behave as a topic-promoting construction.
- Seoane finds that the promoted subject tends to be pragmatically more topic-like than the demoted agent.
  - This is presented as support for the theory that passivization was used to ‘replace’ unaccented topicalization as an information structural strategy.
Problems with the Seoane study

- This study gives us some data about the information structure of passives in the time periods studied.
  - However, it offers no direct comparison with the information structure of topicalization.
  - The study also focuses only on texts written in or after late Middle English.
  - It is, at most, an indirect investigation of the main question.

- In building her study, Seoane also relies on existing assumptions about the properties of topical elements.
  - There is still a great deal of disagreement and inconsistency about the inherent properties of topics, or even the status of topics as a primitive notion (cf. Prince, 1999).
  - We may not wish to structure our inquiries around such pre-theoretical or uncertain assumptions.
Investigating the question

- **Is there a “better” way to approach this question?**
  - Can we investigate the hypothesis without relying on any assumptions about the basic properties of topics?
  - Can we find a way to compare passivization and object topicalization?

- A careful use of corpus data can make this possible.
  - We can quantitatively compare English to Germanic languages which allow unaccented object topicalization (here, Icelandic and German).
  - We can also quantitatively compare historical stages of English.
  - Then, using **parallel parsed corpora**, we can make a more detailed investigation of individual utterances.

- As we will see, this actually informs our understanding of the information structure of these constructions, rather than relying on assumptions about it.
Investigating the question

- What we will find is that the data do not support the intuitively favorable hypothesis.

- A superficial consideration of the corpora suggest that there is a relationship between passivization and unaccented topicalization.

  - But on a more detailed investigation of the corpora, we find that the higher rate of passives in Modern English is not directly related to the loss of unaccented topicalization.
  - Instead, it is associated with completely different (and unexpected) factors.
  - In almost every case, where the English has a passive, clauses in corresponding V2 languages have the same structural subject.

- In this way, the data eventually lead us to an unexpected result.

  - We are led to re-evaluate a hypothesized information structural link, and revise our theoretical assumptions.
The use of parsed corpora in information structural research

Investigating the question

The parsed corpora

- The Icelandic Parsed Historical Corpus (IcePaHC) (Wallenberg et al., 2011)
  - ≈500,000 words at the time of study
- The PPCEME (Kroch, Santorini, and Diertani, 2004)
- The Parsed Corpus of Early New High German (Light, 2011)
  - ≈70,000 words at the time of study
- 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
- We will augment this parallel corpus with a study of three translations of the Rule of St. Benet: Old English (11th c.), Northern Middle English (15th c.), and Southern Middle English (15th c.).
Parallel Parsed Corpus of the New Testament

- Translations of the same text, but not slavish ones.
  - Protestant Bible translations were meant for general public.
  - 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.
Comparing English and its closest relatives

- First, we can quantitatively test the most basic hypothesis: passivization and unaccented topicalization serve the same information structural purpose.
  - For this, we compare data from the three New Testament translations.
- German and Icelandic represent two languages with unaccented object topicalization.
  - We hypothesize that these languages should passivize at a significantly lower rate than English.
The use of parsed corpora in information structural research

Investigating the question

Comparing English and its closest relatives

Quantitative evidence

- In fact, the EME translation of John has a significantly higher frequency of passives than both the ENHG and Icelandic.

<table>
<thead>
<tr>
<th></th>
<th>Passive</th>
<th>Active</th>
<th>Freq. Passive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tyndale</td>
<td>140</td>
<td>1113</td>
<td>0.112</td>
</tr>
<tr>
<td>Luther</td>
<td>101</td>
<td>1262</td>
<td>0.074</td>
</tr>
<tr>
<td>Oddur</td>
<td>81</td>
<td>1236</td>
<td>0.062</td>
</tr>
</tbody>
</table>

EME vs. ENHG: $\chi^2 = 10.569$ on 1df, $p = 0.00115$
EME vs. Icelandic: $\chi^2 = 19.9766$ on 1df, $p \approx 0$
ENHG vs. Icelandic: $\chi^2 = 1.4862$ on 1df, $p = 0.2228$

- These data seem to support the hypothesis that V2 topicalization and passivization are information structurally equivalent.
Another critical element of the hypothesis is the assumption that the rising rate of passives in English is related to the loss of V2 word orders.

For this, we run similar queries to those presented in the cross-linguistic comparison, but now on different historical stages of English.

We hypothesize that varieties of English with higher frequencies of V2 word orders should show lower frequencies of passivization.
Quantitative evidence

We compare multiple Old and Middle English translations of the Rule of St. Benedict.

We compare three prose translations of the Rule of St. Benet: Old English (11th c.), Northern Middle English (1425), Southern/Kentish Middle English (1490).

- PPCME2 (Kroch and Taylor, 2000)
- YCOE (Taylor, Warner, Pintzuk, and Beths 2003)

<table>
<thead>
<tr>
<th></th>
<th>Passive</th>
<th>Active</th>
<th>Freq. Passive</th>
</tr>
</thead>
<tbody>
<tr>
<td>OE Rule</td>
<td>346</td>
<td>1033</td>
<td>0.251</td>
</tr>
<tr>
<td>North ME Prose Rule</td>
<td>214</td>
<td>989</td>
<td>0.178</td>
</tr>
<tr>
<td>William Caxton’s Rule</td>
<td>84</td>
<td>178</td>
<td>0.321</td>
</tr>
</tbody>
</table>

OE vs. North ME: $\chi^2 = 19.7409$ on 1df, $p \approx 0$

OE vs. South ME: $\chi^2 = 5.1774$ on 1df, $p = 0.02288$

OE vs. North ME vs. South ME: $\chi^2 = 34.1658$ on 2df, $p \approx 0$
What do the data tell us?

- The Rule of St. Benedict translations seem to further support the hypothesis.
  - As in the cross-linguistic New Testament comparison, varieties with a V2 grammar have a significantly lower frequency of passivization.
  - Northern Middle English, which has Icelandic-type V2 due to language contact (Kroch and Taylor, 1995; Kroch, Taylor, and Ringe, 1995), has the lowest frequency of passivization of the three.
  - Old English, which has a V2 option due to the low subject position (cf. Haeberli, 1999, 2002, 2005), still has a significantly lower frequency of passivization than the Southern Middle English text.

- However, surprisingly, a more detailed examination of the quantitative evidence suggests that the data in both case studies is misleading.
Pushing the hypothesis

- The data suggest that the presence of V2 options in varieties of English may somehow “lead to” lower frequencies of passivization.

- This leads us to a prediction: if we look at earlier translations of the New Testament, we expect to find lower frequencies of passivization than in the Tyndale.
  - The Wycliffe is a partial sample of the Gospel of John translated into Middle English (PPCME2).
  - The West Saxon Gospels contain an Old English translation of all four Gospels (YCOE).

<table>
<thead>
<tr>
<th></th>
<th>Passive</th>
<th>Active</th>
<th>Freq. Passive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tyndale (EME)</td>
<td>140</td>
<td>1113</td>
<td>0.112</td>
</tr>
<tr>
<td>Wycliffe (MidEng)</td>
<td>72</td>
<td>566</td>
<td>0.113</td>
</tr>
<tr>
<td>West Saxon Gospels (OE)</td>
<td>710</td>
<td>4928</td>
<td>0.126</td>
</tr>
</tbody>
</table>

- In fact, the frequencies of passivization are nearly identical across the three translations.
Considering V2 in the English New Testament texts

Although the frequencies of passivization are nearly identical across the English New Testament translations, they diverge significantly with respect to the presence of V2 word orders.

Consider the rate at which object topicalization triggers subject-verb inversion in each text.

<table>
<thead>
<tr>
<th>Total top.</th>
<th>With inversion</th>
<th>Freq. inversion</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>With full DP subjects:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tyndale</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Wycliffe</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>West Saxon Gospels</td>
<td>15</td>
<td>13</td>
</tr>
<tr>
<td><strong>With pronominal subjects:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tyndale</td>
<td>36</td>
<td>23</td>
</tr>
<tr>
<td>Wycliffe</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>West Saxon Gospels</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>
Considering V2 in the English New Testament texts

- Each of the New Testament translations seems to represent a different level of access to a V2-generating grammar.
  - The West Saxon Gospels have Old English-style V2 orders: subject-verb inversion occurs with full DP subjects (which occupy a low subject position), but not with pronominal subjects (which occupy a high subject position, above Tense).
  - The Wycliffe has almost no object topicalization of any kind, and no subject-verb inversion with topicalized objects.
  - The Tyndale, possibly due to influence from the Luther New Testament, has a high rate of subject-verb inversion with topicalized objects, and in fact allows inversion with all subject types. Seems to have some access to a generalized V2 grammar.

- The different levels of access to V2-generating grammars does not seem to affect these authors’ use of the passive at all.
If it’s not V2, then what’s causing the effect?

- By comparing different translations of a single text, we are given an opportunity to compare the use of passivization not only on a broad quantitative level, but also more directly.
  - Where a passive in one language or variety has been translated as a non-passive elsewhere, we may look more closely to see what syntactic choice was made in place of the passive.
  - This allows us to make a more fine-grained study of how different languages may choose to encode the same informational content.

- A verse-by-verse comparison of the two text groups reveals that clauses with passivization do not generally correspond to clauses with unaccented V2 topicalization.

- Instead, other interesting patterns will become apparent.
Parallel translations of the Rule of St. Benedict

- We considered each attested example of a passive in the Caxton translation of the Rule of St. Benedict, and compared them to the corresponding verses in the Northern Prose translation.
  - Of these 84 tokens, 19 had no corresponding verse, and thus could not be included in the comparison.
  - 15 of the remaining 65 were also passives in the Northern Prose.
  - We compared 50 clauses in which a passive in the Caxton was translated as a non-passive in the Northern Prose Rule.

- The most striking fact is that 13 tokens (26%) involved an instance of impersonal *man* as the subject of a transitive in the Northern Prose Rule, corresponding to a passive in the Caxton.

- In comparison, only 3 (6%) of the passive subjects in the Caxton correspond to a topicalized object in the Northern Prose Rule.
Impersonal *man* in St. Benedict

- Impersonal *man* represents an alternative to the passive which was available to the Northern Middle English and Old English translators, but not to the author of the Caxton translation.

\[(4)\]  
a. *lete them twyes or thries be correct* (Caxton)  
b. *Man sal saie til hir an time, and a-nōðir time, and te þridde.* (Northern Prose Rule)  
*Chapter 33, Verse 7*

- Overall, impersonal *man* was a fairly common option in both the Old English and Northern Middle English translations, but ungrammatical for Caxton.

<table>
<thead>
<tr>
<th></th>
<th>All trans. actives</th>
<th>Thereof with <em>man</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Old Eng.</td>
<td>980</td>
<td>75 (7.7%)</td>
</tr>
<tr>
<td>Northern Prose</td>
<td>889</td>
<td>55 (6.19%)</td>
</tr>
<tr>
<td>Caxton</td>
<td>162</td>
<td>0 (0.0%)</td>
</tr>
</tbody>
</table>
The visible link between passives and impersonal *man* does not hold in the New Testament translations; in fact, no subjects in the Tyndale correspond to impersonal *man* in the Luther, or *maður* in the Oddur.

In a large majority of cases, the translations choose a different clause type, but **preserve the same structural subject**.

- This seems to suggest something about the information-structural properties of the subject position.
- The different translations prefer to express the same entity as subject, regardless of clause type.
Preserving the structural subject

- We perform a verse-by-verse comparison of the three translations.
- In the Luther text, there are 35 non-passive clauses corresponding to passives in Tyndale.
  - Of the 35, 10 (28.6%) use the German verb *werden* (‘become’).
  - An additional 9 (25.7%) are translated as reflexives in the German.
  - 5 (14.3%) correspond to intransitives in the German, and 6 (17.1%) to transitive clauses.
- 33 out of 35 (94.3%) non-corresponding clauses had the same referent as the structural subject in the Luther as in the Tyndale.
Preserving the structural subject

- 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.)

- 72 out of 77 (93.5%) non-corresponding clauses have the same referent as the structural subject in the Oddur as in the Tyndale, including 49/50 of the middles and 8/9 of the actives.
Examples

(5) **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)
Examples

(6) **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. ...en yðar hryggð skal snúast í fögnuð.  
   ...but your sorrow shall turn-MID into joy.  
   (Oddur)
Preserving the structural subject

- Each language uses different syntactic resources to make a certain entity the syntactic subject.

- Although Tyndale and Oddur were strongly influenced by Luther’s translation, this alone cannot account for the effect observed here.
  - In the Wycliffe sample, there are 27 passives which correspond to non-passive clauses in the Luther. Of them, 24 (88.9%) have the same referent as the structural subject.
  - Wycliffe was written at least a century prior to Luther, and thus the close relationship of influence is not in play. However, the effect is still very visible.

- Thus, a more fine-grained consideration of this data does not show a close link between passives and unaccented topicalization.
  - ‘Preserving’ the structural subject seems more important in these translations.
Results of the study

- Initially, corpus data seemed to support the hypothesis that the higher rate of passives in Modern English is related to the loss of V2 word orders.
  - Both German and Icelandic have significantly lower rates of passivization.
- But on more careful investigation, the data seem to be related to independent differences between the languages considered.
  - The priority appears to be preserving the structural subject, regardless of the clause type.
  - We may hypothesize that the subject position has unique information structural properties, which are not identical to those of topicalized objects.
The information structure of the subject position

- Passivization and unaccented topicalization are not being treated as information structurally equivalent; in fact, passive subjects in non-V2 texts are rarely translated as topicalized objects in parallel V2 translations.

- Does this mean that the standard information structural analysis of passivization (and subjecthood) is simply incorrect?
  - No, this does not seem to be the case.

- As a preliminary test of this, we return to the 33 Tyndale passives which correspond to non-passives in the Luther.
  - Each of these tokens was coded both for discourse status (given, evoked, or new) and focus structure (VP focus, narrow focus on a constituent, or focus broader than the VP).
  - We avoid classification of topics in general because an unambiguous classification of topics is rarely possible, but following Vallduvi (1992) we classify all material outside the focus as part of the “Ground”, or topical material.
The information structure of the subject position

- To recap, Los (2009) and others propose that the crucial IS/discourse properties of the subject involved topichood and givenness/referentiality.
- Both claims seem to fit the data.
  - Absolutely no clauses are plausibly analyzed as having narrow focus on the subject, and only in 6 (18.2%) is the focused constituent plausibly broad enough to encompass the subject.
  - This means that 27 (81.8%) of the clauses have the subject as some part of the Ground.
  - 23 (69.7%) of these were unambiguously VP-focused clauses. In these examples, the subject is thus the only constituent in the Ground, and can be assumed to be the topic.
  - Furthermore, 12 (36.4%) of the subjects are referential gaps due to extraction or conjunction, while another 15 (45.5%) are given information. The remaining 6 subjects are evoked or inferrable; absolutely none are discourse-new entities.
The information structure of the subject position

- We conclude that the literature is right about part of the puzzle:
  - The subjects of passives overwhelmingly tend to be both informationally topical and given in the discourse.
  - This fits with the claims made by Seoane (2006).
  - However, they are *not* information structurally equivalent to unaccented V2 topics.

- We do not want to argue against the current literature on the information structure of unaccented topics in V2 Germanic languages, which seem to have these same general properties (cf. Frey, 2006; Light, 2012).

- Our understanding of the information structure of these elements is **not sufficiently precise**.
  - Our current descriptive mechanisms seem to define them as essentially equivalent, but data on the actual usage of these constructions shows that this cannot be true.
The information structure of the subject position

There are two directions in which we may need to refine our understanding of the information structure of such constructions:

1. Our information structural categories may not be fine-grained enough, leading to a failure to identify crucial distinctions in the behavior of certain elements.

2. We may need to refine our understanding of the IS-syntax interface: specifically, how information structural constituency and syntactic constituency may interact, and how this may effect the choice between syntactic constructions.

Both avenues probably require further study, but for the time being we tentatively propose that further consideration of the latter may help us with the problem at hand.
A new hypothesis

- A significant portion of our sample of Tyndale passives had VP focus, leaving the subject as the only constituent in the Ground.
  - Because the corresponding Luther non-passives are encoding the same information, the structural subject is also the Ground in those cases.

- We hypothesize that the subject position may be a preferred position to mark the Ground as a *complete constituent*.
  - That is, when the information structural Ground constituent does not map to multiple syntactic constituents, and when the syntactic constituent does not map to multiple information structural constituents.
A new hypothesis

- Passivization may then be used to syntactically partition the information structural constituents, by moving the entirety of the Ground out of the focused constituent.

- Compare to V2 topicalization, which generally raises only a portion of the Ground; consider cases in which a non-focused adverb may occupy the preverbal position.

(7) Gestern habe ich nur zwei Bücher verkauft! yesterday have I only two books sold
‘I only sold two books yesterday!’
Concluding remarks

- It is not yet clear whether this is the correct analysis for the phenomenon under consideration.
- The main point for today is that the corpus data leads us to **revise our theoretical assumptions** about a relatively central issue in information structural research.
  - The relationship between information structural topics and the structural subject position has been under great debate.
  - We now have reason to believe that more investigation is needed, and some leads to start from.
- We see how corpus data, when appropriately applied, is not simply useful for **supporting our theoretical assumptions**, but also for **refuting them**.
- This puts us on the cutting edge of information structural research.
Structure of the investigation

The corpus research proceeded in the following manner:

1. Queries were constructed to calculate the frequency of passivization in each language/variety in question.
   - First, extract all passive clauses.
   - To calculate the overall frequency of passivization, these were compared to all clauses which could have been passive.

2. Then, we construct queries to calculate the verb-second frequencies in different stages of English.
   - These will be very similar to the sorts of queries Beatrice and Tony discussed this morning.

3. We can plug these numbers into R to calculate $\chi^2$ values.

4. Extracted passive sentences can now be easily accessed in query output, in order to hand-code for information structural information.
Bibliography I


Bibliography II


Bibliography III


Bibliography IV


Bibliography V


