On the use of passives across Germanic

Caitlin Light and Joel Wallenberg

University of Pennsylvania

June 4, 2011

Introduction

- In this paper, we test the hypothesis that deaccented topicalization in a V2 language encodes the same information structure as passivization in a non-V2 language.
  - We use quantitative evidence from parallel parsed corpora in order to explore the use of these syntactic options.
  - A cursory overview of the data appears to support a parallel between V2 topicalization and passivization.
  - However, when the data is examined in more detail, this strong and attractive hypothesis is ultimately falsified.
- This leaves a question: although the two constructions are not the same, it is not obvious how to distinguish between their information structures under current assumptions about IS.
  - Therefore, we will argue that this negative result is most useful in showing where the current understanding of the syntax-information structure interface is lacking.
  - Finally, we suggest a way that the interaction between information structural constituency and syntactic constituency may help to distinguish between these two constructions, and suggest a new line of research into the syntax-IS interface.

Outline

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? Furthermore, however, what remains the same?
- 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* be 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 over 500,000 words (as of last week)
- The Penn Parsed Corpus of Early Modern English (Kroch, Santorini, and Delfs, 2005)
  - ∼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
- 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 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.

Passivization and V2 topicalization as IS equivalents?

- Recent work on the syntax/information structure interface introduces the proposal that deaccented 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 geseet?
Lord have-you not good seeds on your acre sowed
wo her hatt er denn das vnkraut? vnd er sprach, **das**
where from has he then the weeds and he spoke this
hat eyn feyndt than
has an enemy done

b. This was done by an Enemy.
(our constructed example)

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, deaccented position (cf. Los, 2009; Seoane, 2006).
- These proposals argue that both deaccented topicalization and promotion of an argument in the passive result in marking the DP as informationally topical/thematic, and/or as a discourse given entity.
- This argument has intuitive appeal, and the information structural/pragmatic claims have much support in the general literature on these constructions.
- We will quantitatively test this hypothesis, by considering whether the frequency of passives seems to be affected by the presence of a V2 grammar.

Quantitative evidence

- A quantitative comparison of the parallel New Testament samples shows that the EME translation of John has a significantly higher frequency of passives than both the ENHG and Icelandic.

	Passive	Active	Freq. Passive
Tyndale	140	1113	0.112
Luther	101	1262	0.074
Oddur	81	1236	0.062

EME vs. ENHG: Chi-square = 10.569 on 1df, $p = 0.00115$
EME vs. Icelandic: Chi-square = 19.9766 on 1df, $p \approx 0$
ENHG vs. Icelandic: Chi-square = 1.4862 on 1df, $p = 0.2228$

- These frequencies seem to support the hypothesis that V2 topicalization and passivization are information structurally equivalent, because the Germanic languages with V2 are shown to passivize at a lower rate.

Quantitative evidence

- A second test case can be found by comparing 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).
  - Penn Parsed Corpus of Middle English (Kroch and Taylor, 2000)
  - York Corpus of Old English Prose (Taylor, Warner, Pintzuk, and Beths 2003)

	Passive	Active	Freq. Passive
OE Rule	346	1033	0.251
North ME Prose Rule	214	989	0.178
William Caxton's Rule	84	178	0.321

OE vs. North ME: Chi-square = 19.7409 on 1df, $p \approx 0$
OE vs. South ME: Chi-square = 5.1774 on 1df, $p = 0.02288$
OE vs. North ME vs. South ME: Chi-square = 34.1658 on 2df, $p \approx 0$

What does this data tell us?

- The Rule of St. Benedict translations seem to further support the hypothesis that deaccented V2 topicalization and passivization are information structurally equivalent.
  - 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.
- We will argue that the differences in frequencies of passivization across these varieties is, in fact, independent of the presence of a V2 grammar.

Pushing the hypothesis

- The data from the Rule of St. Benedict case study suggests 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).

	Passive	Active	Freq. Passive
Tyndale (EME)	140	1113	0.112
Wycliffe (MidEng)	72	566	0.113
West Saxon Gospels (OE)	710	4928	0.126

- 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, however, these texts diverge significantly with respect to the presence of V2 word orders.
  - To demonstrate this, we calculated the rate at which object topicalization triggers subject-verb inversion in each text.

	Total top.	With inversion	Freq. inversion
With full DP subjects:			
Tyndale	8	8	1.000
Wycliffe	0	0	N/A
West Saxon Gospels	15	13	0.867
With pronominal subjects:			
Tyndale	36	23	0.639
Wycliffe	5	0	0.000
West Saxon Gospels	4	0	0.000

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, as expected, 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. This text seems to demonstrate little or no remnants of the declining Old English V2 grammar.
  - The Tyndale, quite 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. Tyndale seems to have some access to a generalized V2 grammar.
- However, 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.
- We will show that a verse-by-verse comparison of our two text groups reveals that clauses with passivization do not generally correspond to clauses with deaccented V2 topicalization.
- However, other interesting patterns will become apparent.

If it's not V2, then what's causing the effect? (Rule of St. Benet)

- 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 translation.
  - We were thus able to compare 50 clauses in which a passive in the Caxton was translated as a non-passive in the Northern Prose Rule.
- The most striking fact about this data 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.

(2) a. lete them twyes or thries **be correct** (Caxton)

b. **Man** sal saie til hir an time, and a-noþir time, and te þridde. (Northern Prose Rule)
(Chapter 33, Verse 7)

- Considering the overall frequency of *man* occurring as a subject in active transitive clauses shows that it was a fairly common option in both the Old English and Northern Middle English translations, but ungrammatical for Caxton.

	All trans. actives	Thereof with *man*
Old Eng.	980	75 (7.7%)
Northern Prose	889	55 (6.19%)
Caxton	162	0 (0.0%)

If it's not V2, then what's causing the effect? (New Testament)

- 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.
- What we find instead is that, 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.
  - In order to express the same informational content, these different translations apparently prefer to express the same entity as subject, regardless of clause type.

Preserving the structural subject

- In the Luther text, there are 35 non-passive clauses corresponding to passives in Tyndale.
  - Of the 35, 10 (28.6%) involve the use of the German verb *werden* meaning '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 transitives 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

(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)

Examples

(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. . . . 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.

The information structure of the subject position

- The data shows that in the texts under consideration, passivization and deaccented 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. This allows us to examine the information-structural properties of a small sample set.
  - 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 Vallduví (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 therefore 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.
  - However, they are *not* information structurally equivalent to deaccented V2 topics.
- We also do not want to argue against the current literature on the information structure of deaccented topics in V2 Germanic languages, which seem to have these same general properties (cf. Frey, 2006).
- Instead, we propose that our data be taken as evidence that 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 set 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.

(5) Gestern habe ich nur zwei Bücher verkauft!
yesterday have I only two books sold
'I only sold *two books* yesterday!'

Conclusion

- In this paper, we discussed both theoretical and quantitative evidence suggesting that deaccented V2 topicalization and passivization are information structurally equivalent.
- We showed, based on evidence from a comparative study across several Germanic languages and several stages in the history of English, that a detailed study of the data cannot support such a claim.
  - Using two separate case studies, we identified independent causes of the disparate frequencies of passivization found in the relevant texts.
  - Furthermore, we showed that authors in a parallel corpus of these languages do not use passives and V2 topicalization in the same contexts.
- However, it does not seem to be the case that the existing information structural analyses of either construction are actually false.
  - Instead, we proposed a refinement of our analysis of these elements, by positing how interactions between information structural and syntactic constituency may affect the choice of how information is syntactically represented.
  - We take this as a working hypothesis in reaction to the observed results, and in future work we hope to test it further.

Acknowledgements

We would like to thank Tony Kroch for his help, and all the attenders of SHES 9 for their discussion on an earlier version of this work.

Bibliography I

- Frey, W. (2006). "Contrast and movement to the German prefield". In: *The architecture of focus* 235264.
- Haeberli, Eric (1999). "On the word order 'XP-subject' in the Germanic languages". In: *Journal of Comparative German Linguistics* 3.1, pp. 1–36.
- — (2002). *Features, Categories and the Syntax of A-Positions: Cross-Linguistic Variation in the Germanic Languages*. Studies in Natural Language and Linguistic Theory 54. Dordrecht: Kluwer Academic Publishers.
- — (2005). "Clause type asymmetries in Old English and the syntax of verb movement". In: *Grammaticalization and Parametric Change*. Ed. by M. Batllori and F. Roca. Oxford: Oxford University Press, pp. 267–283.
- Kroch, Anthony, Beatrice Santorini, and Lauren Delfs (2005). "Penn-Helsinki parsed corpus of Early Modern English". Size 1.8 Million Words.

Bibliography II

- Kroch, Anthony S. and Ann Taylor (1995). "Verb Movement in Old and Middle English: Dialect Variation and Language Contact". In: *Parameters of Morphosyntactic Change*. Ed. by Ans van Kemenade and Nigel Vincent. Cambridge: Cambridge University Press.
- — (2000). "Penn-Helsinki Parsed Corpus of Middle English. CD-ROM. Second Edition." Size: 1.3 million words.
- Kroch, Anthony S., Ann Taylor, and Don Ringe (1995). "The Middle English verb-second constraint: a case study in language contact and language change". In: *Textual Parameters in Older Language*. Ed. by Susan Herring et al et al. John Benjamins.
- Light, Caitlin (2011). "Parsed Corpus of Early New High German". URL `http://enhgcorpus.wikispaces.com/`.
- Los, B. (2002). "The loss of the indefinite pronoun man". In: *Selected papers from 11 ICEHL: Santiago de Compostela, 7-11 September 2000*, p. 181.

Bibliography III

- Los, B. (2009). "The consequences of the loss of verb-second in English: information structure and syntax in interaction". In: *English Language and Linguistics* 13.01, pp. 97–125.
- Seoane, E. (2006). "Information structure and word order change: The passive as an information-rearranging strategy in the history of English". In:
- Taylor, Ann et al. (2003). "The York-Toronto-Helsinki Parsed Corpus of Old English Prose".
- Vallduví, Enric (1992). "The Informational Component". PhD thesis. Philadelphia: University of Pennsylvania.
- Wallenberg, J. C. et al. (2011). "Icelandic Parsed Historical Corpus (IcePaHC)." Version 0.4. Size: 440 thousand words. URL `http://www.linguist.is/icelandic_treebank`.