# New changes in English

## A diachronic perspective on the relation between newness and syntax*

Erwin R. Komen
Radboud University Nijmegen

### 1. Introduction

In this paper, I will address the mapping between the changing syntax of English, and the expression of the information status categories "given" and "new".

Present-day English favours clause-initial subjects that are linked to the preceding discourse, while it tends to have clause-final objects convey discourse-new information (one of the first to discuss this was Halliday 1967, a more recent study is Prince 1992). This has not always been the case. In Old English, new subjects could precede discourse-old information, as illustrated by example (1a).[1] The subject, *stilnes and swige* 'silence and quietness', is the new information in the sentence. Later translations of this example use various strategies in order to prevent discourse new subjects. The Late Modern English translation in (1b) uses an expletive subject, so that the logical subject becomes a predicative noun phrase, which is acceptable as new. A translation into Present-day English is (1c), where the inanimate 'hall' is the subject and what used to be the subject in OE is a subject complement. The example illustrates that the position of given and new information has changed too. Where the Old English original has New-Given, the Present-day English translation has Given-New.

(1) a. Ða weard stilnes and swige geworden innon đare healle.
then was silence and quiet become inside that hall
(coapollo, ApT:16.30.332) [2]
b. 'Then there was stillness and silence within the hall.' (Thorpe 1834)
c. 'Then the hall had become still and quiet.'[4]

By investigating this diachronic change in the mapping between information status and syntax, we stand to learn more about how information structure and syntax interact in general.

*Linguistics in the Netherlands 2011*, 74–85. DOI 10.1075/avt.28.07kom
ISSN 0929–7332 / E-ISSN 1569–9919 

Old English has been regarded as a V2 language, allowing its first position to host constituents from different syntactic categories (such as subject, object, adjunct).[4] The English V2 rule was lost in the 15th century, which, together with a change from OV to VO word order, resulted in its current SVO syntax (Kemenade & Westergaard forthcoming, Warner 2007). The loss of V2 meant that the default position for subjects became the preverbal one. The combination of the syntactic change and the principle that Given tends to precede New information (Gundel, 1988), should, at first glance, have led to an increase in subjects conveying Given information, and, the flipside of the coin, in non-subjects (direct object, indirect objects, PP adjuncts) conveying New information. But the Given-before-New principle is not the only information ordering effect that is at work in English. Los & Komen (forthcoming) have shown that this principle is more often violated in earlier periods of English than in Present-day English, because contrastive new information may be positioned in the first position of the clause in earlier stages of English. In that respect, Old English is like Present-day Dutch and German (Rinke & Meisel 2009). New information, then, can be expected to be more often expressed preverbally in Old English, especially if this information is contrastive.

This leads me to posit the following two hypotheses for the diachronic change in the syntactic expression of newness in English:

(2) **The position of New information**
The percentage of New information following the finite verb (or auxiliary) in English increases over time.

(3) **The grammatical category of New information**
The percentage of non-subjects expressing New information increases over time.

This paper presents a corpus-based approach to verify these hypotheses by means of a pilot study that makes use of a syntactically annotated corpus that is enriched with referential information. Section 2 defines the notions of newness used in this research, and the corpus-based approach is described in Section 3. The results of the pilot are presented in Section 4, and the last section discusses the implications of the findings.

## 2. What is new?

### 2.1 Defining new

Before we can look into the behaviour of constituents expressing new information, we need a working definition of the information status "New". We define "New" referentially: a constituent is referentially new when it is not related to anything else in the preceding text or the wider context. In other words — it is not yet present in the *common ground* between speakers (Lambrecht 1994: 59; Stalnaker 1974). Only referentially new constituents should be regarded as "New" for information ordering purposes, since speakers order their information according to their assumptions about what information their interlocutor already has (i.e. the common ground). The examples in (4) illustrate which types of discourse-new elements are not necessarily new in a referential sense.

(4) a. Once upon [$_{\text{New}}$ a time] there was [$_{\text{New}}$ a linguist].
b. The Romans killed 20,000 of the rebels. [$_{\text{Inferred}}$ The rest] escaped.
c. Mary went to [$_{\text{New}}$ [$_{\text{Identity}}$ her] room].
d. I was looking at [$_{\text{Assumed}}$ the moon] last night.
e. Susan entered the room. [$_{\text{Identity}}$ She] sat down.
f. Linguists are [$_{\text{Inert}}$ servants of [$_{\text{Inert}}$ mankind]].
g. The students turned to [$_{\text{Inert}}$ robbery].

Of (4a–g), only *a time* and *a linguist* in (4a) are wholly new. The identity of *The rest* in (4b), though not mentioned before, can be "Inferred" from the preceding discourse. *Her room* in (4c) links to *Mary* through the anchor *her*; *moon* in (4d) does not connect to the discourse, but is "Assumed" to be known between interlocutors; *she* in (4e) has an "Identical" referent established in the discourse. Then there are elements that are "Inert" from a referential point of view, since they appear in syntactic functions that disqualify them from establishing a referent that can be linked back to, like *servants*, *mankind* and *robbery* in (4f,g). These findings give us a working definition of referential newness as in (5), and a number of distinct referential categories ("Inferred", "Assumed" and "Identity") that help us to define a hierarchy of newness in (6).

(5) *Definition of Referentially New*
A constituent is referentially new if it refers to a referent that has not been mentioned in prior discourse, is not assumed to be known by the hearer, does not contain an anchor to an established referent, and can be referred back to in subsequent clauses.[5]

## 2.2 Relative newness

The hypothesis in (2) requires a way to identify which items in a clause are newer than other items, and hence a hierarchy that orders referential states like "Inferred", "Assumed" and "Identity" on a scale. For the purposes of the present investigation, I have set up a tentative hierarchy of decreasing newness as in (6).

(6) **Newness hierarchy**
New > Inferred > Assumed > $\text{Identity}_{\text{non-salient}}$ > $\text{Identity}_{\text{salient}}$

This hierarchy allows us to determine when one constituent is relatively newer than another. The newest information is "New" in the sense that it does not relate to anything in the common ground. Next comes "Inferred" information, which relates to something in the common ground, but still establishes a discourse-new referent. "Assumed" information links to the common ground, though not through an overt antecedent. Next is information of the type "Identity", which establishes a direct referential relationship with an entity in the common ground. Non-salient elements (that is, elements with a relatively distant antecedent) in the common ground are, of course, relatively newer than salient ones (those with a relatively nearby antecedent).[6]

The next section will show how constituents in existing texts can be annotated for the basic information status types, and how these corpora can subsequently be searched, in order to verify the hypothesis stated in the introduction.

## 3. Looking for new information

The corpus-based investigation into newness described in this paper is done in two stages. The first stage, described in Section 3.1, is to label the relevant constituents with information status primitives like "Inferred", "Assumed" and "Identity". The second stage, described in Section 3.2, queries the available texts in order to see how the relation between newness and syntax has changed.

### 3.1 Coreference resolution

The investigation can make use of four corpora containing texts from Old English (starting about A.D. 700) to Present-day English, totalling 6 million words (Kroch et al. 2004; Kroch et al. 2010; Kroch and Taylor 2000; Taylor et al. 2003). They comprise various genres and are syntactically annotated using a bracketed labelling Treebank format (Marcus et al. 1994).

The syntactic annotation of the corpora establishes the phrasal category of constituents, their grammatical function, and their structural position in the clause. It does not, however, provide the information status for the constituents. The process of labelling all relevant constituents in a text for information status is known as "coreference resolution" in computational linguistics (see, for example, Soon et al. 2001). Coreference information shows for each noun phrase whether it refers back to another constituent, and, if so, with what type of link ("Inferred", "Assumed" or "Identity"). The result of coreference resolution is, perhaps, not completely what we are looking for. Coreference resolution gives us the categories "Identity" and "Inferred", but we need a finer-grained distinction between the non-referring types "New", "Assumed" and "Inert", in order to verify the hypotheses in the introduction.

For the purpose of labelling all noun phrases with information status primitives, I have constructed a computer program called "Cesax" (Komen 2011). This program resolves coreference semi-automatically: it does what it can automatically and asks for user input in ambiguous cases.[7]

Texts that are enriched with coreference information using Cesax follow an XML standard described by Komen (2011), and can be queried using the Xquery language (Boag et al. 2010). The next section describes the algorithms necessary to find referentially new constituents (see definition 5) and constituents that are relatively the newest in a clause. The output of these algorithms should be combined with syntactic information in order to verify the hypotheses (2) and (3) in the introduction.

### 3.2 How to look for new information

Referentially new constituents can be recognized in a text by checking two things. The first step is to check if an NP has been annotated with the information status "New". If that is so, then the NP as a whole can be regarded as referentially new, since it does not have an information status "Assumed" or "Inert", because if it had, it would have received that label. The second step is to make sure the constituent does *not* contain an anchor, an element of a noun phrase that links back (see 2.1).

If we want to check whether the position where the *newest* constituent in the clause occurs has changed over time, we need to use a different algorithm. The relatively newest constituent does not necessarily have to be marked as being *absolutely* "New". The algorithm to check for relative newness considers the information status categories of all relevant phrases in a clause, and then employs the Newness hierarchy defined in (6) above to determine the winning candidate.

(7) *Algorithm to get the relatively newest NP in a clause*
Step 1: Get a list of all the NPs and PPs at the clause-level.
Step 2: Delete empty NPs from the list as well as those marked "Inert".
Step 3: Get the number of NPs marked "New" that are unanchored:
  If there is 1, this is the relatively newest;
  If there is more than 1, return the syntactically most prominent;[8]
  If there are 0, continue with step 4.
    Step 4: Repeat step 3 for NPs marked "Assumed".

Step 5: Get the number of remaining NPs (these are either directly referential, or they have a referential anchor):
If there is 1, this is the relatively newest; Otherwise return the one with the largest antecedent distance.

The procedure to check whether a constituent is referentially new and the one that retrieves the relatively newest NP in a clause have been coded into Xquery as functions *IsNew* and *Newest* respectively. The pilots discussed in the next section make use of these functions, in order to retrieve clauses that fulfil syntactic constraints (like grammatical role and position in the clause) as well as information-structural constraints (like being referentially new or being the relatively newest constituent).

## 4. The relation between newness and syntax

This section describes two pilot studies aimed at verifying the hypotheses (2) and (3) in the introduction, which predict particular aspects of the relationship between syntax and information status. The pilots make use of the first six texts that our research group has enriched with coreference information using Cesax, as listed in Table 1.

**Table 1.** Texts used in the pilot studies

| Name | Words | Period |
|---|---|---|
| Apollonius of Tyre | 6545 | OE (950–1050) |
| Saint Vincent | 728 | OE (1050–1150) |
| Oroonoko | 5475 | eModE (1668–1688) |
| Brightland | 1341 | LmodE (1711) |
| Defoe | 9378 | LmodE (1719) |
| Long | 8851 | LmodE (1866) |

### 4.1 The relation between grammatical role and referentially new

The first pilot is aimed at verifying hypothesis (3), which says that the number of non-subjects expressing referentially new information increases over time. Figure 1 shows the percentage of referentially new subjects, objects and prepositional phrases. The data confirm the main prediction: the percentage of referentially new non-subjects (PPs and NP objects) rises in the course of time.

The data also show that the percentage of referentially new subjects increases over time, doubling from OE to eModE. More data are needed, however, since this change is currently just below the level of significance ($p$=0.052, Fisher's exact test). The reason for this rise (if it does prove to be significant) cannot be found in a difference between transitive and intransitive clauses, because clauses with an object, labelled "with competition" in Figure 1, and those without (labelled "no competition") behave alike. Some possible reasons for the rise of subjects are suggested in Los & Dreschler (forthcoming). For the purposes of this investigation, what is striking in Figure 1 is the change in the information status of referentially new *objects*: in OE, a far higher percentage of them is Given than in later periods. One of the reasons could be the fact that Given objects as first constituents, as in (8), become increasingly disallowed, witness the PDE translation of this example.

(8) (When he saw that these places were locked, he said to a boy: "So be thou in health, tell me for what reasons this city continueth in so great lament and wail?")
Him andswerode se cnapa and þus cwæđ:
him answered the boy and thus said
'The boy answered him and thus said: …' (coapollo,ApT:7.12.100–102)

Examples like (8) disappear as the result of the loss of V2 and the emergence of SVO as the canonical order. The confirmation of hypothesis (3) leads us to argue that the Given-before-New principle is present in OE as well as in PDE. It is the combination of this principle with the syntactic change from OE to PDE — the rise in preverbal subjects — that results in the change in mapping between grammatical category and information status. All this is *not* to say that the frequency of given information preceding new information remains constant in English, since

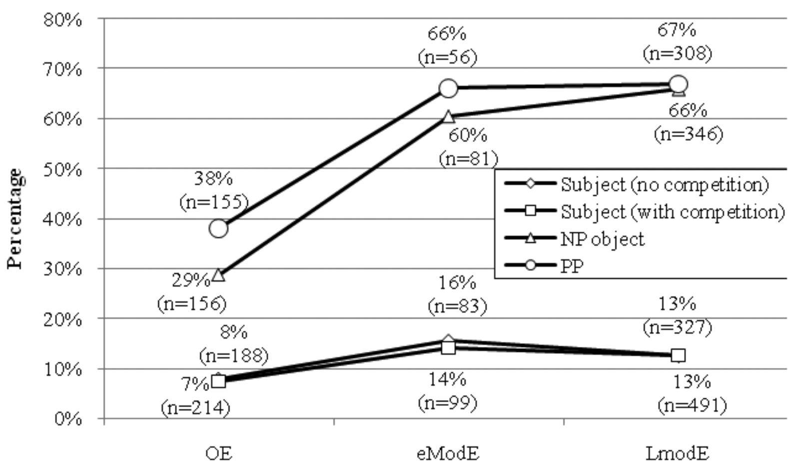


**Figure 1.** Change in the referential newness of constituents over time

there is another tendency in OE (comparable to what is observed in Dutch and German), allowing (contrastive) new objects in the first position.

## 4.2 The syntactic position of the relatively newest constituent

The hypothesis defined in (2) combines two tendencies. The first is that the Given-before-New principle can be seen to operate throughout the history of the English language. We saw this confirmed by the results of the previous section. The second tendency is the change in the contrastive information position from preverbal in OE to postverbal by ME, as explained in Los & Komen (forthcoming). The combination of these two tendencies suggests that the percentage of clauses with the newest information postverbally should increase over time. Testing this hypothesis requires an investigation that combines information status, syntactic function, and position in the clause.

This second pilot finds all clauses in which the relatively newest constituent (as determined by the algorithm in Section 3.2) is in postverbal position, where "postverbal" means that the constituent follows the finite verb or auxiliary.[9] The results in Table 2 confirm the expectations: the percentage of clauses with the relatively newest constituent postverbally increases over time — at least for those clauses where there are two or more NPs that can be ranked according to their relative newness.[10]

**Table 2.** The percentage of newest constituents occurring postverbally

| Period | Clauses with one or more NPs or PPs | Subject only clauses |
|---|---|---|
| OE | 69% (n=214) | 32% (n=270) |
| eModE | 83% (n=98) | 10% (n=110) |
| LmodE | 84% (n=488) | 7% (n=452) |

The data show that there are relatively fewer clauses with the newest constituent postverbally (i.e. after the finite verb or auxiliary) in OE. The most likely explanation is that contrastive, new objects could appear preverbally while English was still a V2 language. But since the annotation system does not mark contrast, this is not something we can measure using an algorithm. I have checked the ten instances where the newest constituent occurred preverbally in the oldest OE text, and found that at least half of them were, in fact, contrastive. Their newest constituents were left-dislocated, and contrasted with one another in the context, as shown by example (9a–b).

(9) a. [Ðe þe his sawle lufæđ], he forlosæđ heo witodlice;
He who his soul loves he looses her in.fact
'Who loves his life will in fact lose it' (covinceb,[Vincent]:290.6)

b. and [þe đe his sawlæ hatæđ] on þissere weorulde,
and he that his soul hates in this world
he healt hire sođlice on þam ecan life.
He keeps it in.fact in the eternal life
'and who hates his life here, will in fact keep it in eternal life.'
(covinceb,[Vincent]:290.7)

The general picture, then, confirms hypothesis (2) stated in the introduction.

## 5. Discussion and conclusions

In order to understand the interplay between information status and syntax in general, this paper focuses on a specific case: the diachronic change in the relation between newness and syntax. Combining the loss of V2 in English with the tendency for given information to precede new information, my first hypothesis states that the percentage of non-subjects conveying new information increases over time. The given-before-new tendency, combined with observations about the change in position of contrastive information from other research, leads to a second hypothesis, which predicts an increase in the postverbal realization of the relatively newest information.

The approach to verify these hypotheses is based on annotated corpora. Noun phrases from existing syntactically annotated corpora receive a label indicating their information status. The six texts that have been enriched with coreference information have been used in two pilots. Both pilots confirm the hypotheses that were made. This is a promising result, and an inspiration to continue enriching the English corpora with coreference information, so that these conclusions, necessarily tentative because of the limited size of the database, can be verified by future research.

## Notes

* I would like to thank Bettelou Los, Ans van Kemenade, Rosanne Hebing, Gea Dreschler, as well as the participants of Radboud University's Language in Time and Space workshop for valuable comments. I would like to acknowledge the support of the Netherlands Organization for Scientific Research (NWO), grant 360-70-370.

**1.** The abbreviations used for the different English time periods are as follows: Old English is "OE" (450–1066), Middle English is "ME" (1066–1500), early Modern English is "eModE" (1500–1770), and late Modern English is "LmodE" (1770–1910).

**2.** The references in these examples follow the system of short titles as used in the original Helsinki corpora (Kytö 1993).

**3.** The translation of the examples are the author's, unless indicated otherwise.

**4.** The V2 status of English has been debated by many (e.g.: Speyer 2010).

**5.** I leave the definition of "prior discourse" vague at the moment. I haven't heard of an objective measure yet to establish how far back a constituent can link and still be "given". Centering, for instance, takes a frame of one phrase, while others are using a frame of more phrases. More work needs to be done in this area.

**6.** The newness hierarchy is, for now, but a working hierarchy. It can reversely be compared with other hierarchies, such as Prince's "familiarity scale" (1981: 245), and Lambrecht's "topic acceptability scale" (1994, pp. 165, 262). Both authors divide "new" into "brand new anchored" and "brand new unanchored", and they don't distinguish between more or less salient "identity" entities.

**7.** The actual program adds two more types: CrossSpeech, which is like "Identity", but then either the noun phrase or its antecedent are part of (in)direct speech, and NewVar, which is an empty category needed to identify variables set up, for instance, by *wh*-phrases.

**8.** When there are more NPs marked as "New", then the newest one is the constituent with the grammatical role that is lowest on a scale resembling the accessibility hierarchy as defined by Keenan & Comrie (1977):

Subject > Argument; Possessive > PP-object > Other

**9.** An anonymous reviewer rightly argues that the position of the constituent under consideration with respect to the non-finite verb (if present) should also be taken into account. We intend to do this in future research.

**10.** Due to the algorithm that determines the relatively newest constituent, the subject-only clauses show a decrease in having postverbal constituents that are referentially "newest". Many of these clauses, especially those in LmodE, contain expletives, which are referentially "Inert", and therefore don't join the competition on the newness scale. While such constructions should probably be treated separately, it is still interesting to see that there is a tendency for subject-only clauses to have the subject decreasingly in a postverbal position.

## References

Boag, Scott, Don Chamberlin, Mary F. Fernández, Daniela Florescu, Jonathan Robie & Jérôme Siméon. 2010. *XQuery 1.0: An XML Query Language (Second Edition)*: W3C Recommendation, <http://www.w3.org/XML/Query/#specs>.

Gundel, Jeanette. 1988. *The role of topic and comment in linguistic theory*. New York: Garland publishing company.

Halliday, Michael A.K. 1967. "Notes on transitivity and theme in English, part II". *Journal of Linguistics* 3.199–244.

Keenan, Edward L. & Bernard Comrie. 1977. "Noun Phrase Accessibility and Universal Grammar". *Linguistic Inquiry* 8.63–99.

Kemenade, Ans van & Marit Westergaard (forthc.) "Syntax and information structure: V2 variation in Middle English". *Information structure and syntactic change in the history of English*, ed. by Bettelou Los, María José López-Couso and Anneli Meurman-Solin. Oxford: Oxford University Press.

Komen, Erwin R. 2011. Cesax: Semi-automatic coreference resolution. Ms., Radboud University Nijmegen.

Kroch, Anthony, Beatrice Santorini & Ariel Diertani. 2004. *Penn-Helsinki Parsed Corpus of Early Modern English*, <http://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-2/index.html>.

Kroch, Anthony, Beatrice Santorini & Ariel Diertani. 2010. *Penn Parsed Corpus of Modern British English*, <http://www.ling.upenn.edu/hist-corpora/PPCMBE-RELEASE-1/index.html>.

Kroch, Anthony & Ann Taylor. 2000. *Penn-Helsinki Parsed Corpus of Middle English, second edition*, <http://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-2/>.

Kytö, Merja. 1993. *Manual to the Diachronic Part of the Helsinki Corpus of English Texts: Coding conventions and lists of source texts*. Helsinki: University of Helsinki, English Department.

Lambrecht, Knud. 1994. *Information structure and sentence form. Topic, focus and the mental representations of discourse referents*. Cambridge: Cambridge University Press.

Los, Bettelou & Gea Dreschler (forthc.) "The loss of local anchoring: From adverbial local anchors to permissive subjects". *Rethinking approaches to the history of English*, ed. by Terttu Nevalainen and Elizabeth Closs Traugott. New York: Oxford University Press.

Los, Bettelou & Erwin R. Komen (forthc.) "Clefts as resolution strategies after the loss of a multifunctional first position". *Rethinking approaches to the history of English*, ed. by Terttu Nevalainen and Elizabeth Closs Traugott. New York: Oxford University Press.

Marcus, Mitchell, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz & Britta Shashberger. 1994. "The Penn treebank: Annotating predicate argument structure". Paper presented at *The human language technology workshop, March 1994*, San Francisco, CA.

Prince, Ellen. 1981. "Toward a taxonomy of given-new information". *Radical pragmatics*, ed. by P. Cole, 223–255. New York: Academic Press.

Prince, Ellen. 1992. "The ZPG letter: Subjects, definiteness and information-status". *Discourse description: Diverse analyses of a fund raising text*, ed. by William C. Mann and Sandra A. Thompson, 295–325. Amsterdam/Philadelphia: John Benjamins.

Rinke, Esther & Jürgen M. Meisel. 2009. "Subject-inversion in Old French: Syntax and information structure". *Proceedings of the Workshop "Null-subjects, expletives, and locatives in Romance". Arbeitspapier 123*, ed. by Georg A. Kaiser and Eva-Maria Remberger, 93–130. Konstanz: Fachbereich Sprachwissenschaft, Universität Konstanz.

Soon, Wee Meng, Hwee Tou Ng & Daniel Chung Yong Lim. 2001. "A machine learning approach to coreference resolution of noun phrases". *Computational Linguistics* 27.521–544.

Speyer, Augustin. 2010. *Topicalization and stress clash avoidance in the history of English*. Berlin/New York: Mouton de Gruyter.

Stalnaker, Robert. 1974. "Pragmatic presuppositions". *Semantics and philosophy*, ed. by K. Munitz Milton and Peter Unger, 197–213. New York: University Press.

Taylor, Ann, Anthony Warner, Susan Pintzuk & Frank Beths. 2003. The York-Toronto-Helsinki Parsed Corpus of Old English Prose. In *Electronic texts and manuals available from the Oxford Text Archive*.
Warner, Anthony. 2007. "Parameters of variation between verb-subject and subject-verb order in late Middle English". *English Language and Linguistics* 11.81–111.

### *Author's address*

Erwin R. Komen
Radboud University Nijmegen
Centre for Language Studies
Box 9103, 6500 HD Nijmegen

E.Komen@Let.ru.nl