Quantifying information structure change in English

Quantifying information structure change in English[1]

Erwin R. Komen, Rosanne Hebing, Ans van Kemenade, Bettelou Los

Radboud University Nijmegen

Abstract

The verb-second constraint in Old and Middle English made available a special clause-initial position that could host more than just the subject. Los (2009) suggests that this position served a discourse-linking function, expressed by, for instance, an adverbial. This allowed the subject to be reserved for human "protagonists". It stands to reason that the loss of verb second in the fifteenth century entailed a decrease in the prevalence of discourse-linking clause-initial adverbials. The subject took over the discourse-linking function, thus extending its functional load.

This article tests four hypotheses concerning the changing functional load of the English subject. Our corpus consists of syntactically-parsed texts that have been enriched with referential information, allowing us to quantify the changes affecting the subject.

[1] We would like to acknowledge the support of the Netherlands Organization for Scientific Research (NWO), grant 360-70-370.

1. Introduction

1.1. Old English V2 syntax and the subject

OE resembles ModG and Dutch in the sense that all three are verb-second languages. However, there is an important difference between the OE version of V2 on the one hand and the ModG and Dutch version of V2 on the other: OE V2 allows for two distinct types of verb-movement, yielding either V2 or verb-third surface word order. When the first constituent contains a *wh*-phrase, negation or the narrative foregrounder *þa* 'then', the finite verb moves to the higher position (C) and categorically appears in second position, followed by the subject in third position, irrespective of the form (full NP or pronoun) of that subject, see e.g. Fischer et. al (2000). This is shown in (1), where the subject appears postverbally, whether it is nominal as *seo eadiga Margareta* in (1a) or pronominal as *he* in (1b). This type of verb-movement survives as I-to-C movement (subject-auxiliary inversion) in PDE (Fischer et al. 2000).

(1) a. ða *geherde seo eadiga Margareta* and hi hit on
then heard the blessed Margaret and she it in

bocum fand, þæt þa cinges and þa ealdormennand þa
books found that the kings and the aldermen and the
yfela gerefan ofslogen æfre and bebyrodon ealle þa godes
evil reeves killed ever and buried all the god's
theowas, þe þær on lande wæron. [comargaC.o34:33]
servants who there in land were

'Then the blessed Margaret heard said, and found it written in books, that the kings and aldermen and the evil reeves were constantly killing and burying all the servants of God who were there in that country' (Los 2012)

b. ða he on hiswege rad, þa *beseah he* on þæt eadigan
when he on his way rode, then looked he on that blessed
mæden, þær þe hi sæt wlitig and fægeronmang
maiden there where she sat beautiful and fair among
hire geferan. ða *cwæð he* to hiscnihtum:Ridað hraþe
her companions then said he to his servants ride quickly
to þære fæmnan and axiað hire, gifhi seo frig.
to that girl and ask her if she is free
[comargaC.o34:48]

'When he was riding on his way, he beheld that blessed maiden where she was sitting among her companions, beautiful and fair; then he said to his servants: "Ride quickly to that girl and ask her if she is free."' (Los 2012)

The lower verb position in OE main clauses is in evidence when the first constituent is not a *wh*-phrase, negation or *þa* 'then', but an adverbial, such as a topicalized PP, or an object. When the subject of such a sentence is pronominal, it typically follows the clause-initial element and appears in second position – i.e. preverbally –, yielding verb-third word order. This is illustrated in (2), with the finite verb *gelefa* appearing after the pronominal subject *ic*. However, when the subject is nominal, it will follow the finite verb, as witnessed by (3), in which the nominal subject *iosep* follows the finite verb *wæs* (van Kemenade 1987). It should be noted that there are exceptions: some XPs occurring clause-initially – e.g. *witodlice* and *soðlice* – never yield inversion. The same goes for discourse-old nominals. Hinterhölzl and Petrova (2010) argue that in Old High German the type of verb-movement illustrated in (2) and (3) originally served to distinguish topical or given information from new information: the area preceding the finite verb in a sentence like (2) contains the clause-initial adverbial that constitutes a discourse link and a pronominal subject that encodes the protagonist – both given information (see also Los 2012).

(2) Andseo eadiga Margareta hire handan upp ahof and
And the blessed Margaret her hands up lifted and

hi to gode gebæd and þus cwæð:

her to God prayed and thus spoke:

On þe ***ic*** *gelefa* leofa Drihten, [comargaC.o34:113-116]

On thee I believe dear Lord

'And the blessed Margaret lifted up her hands and prayed to God and spoke thus: "In you I believe..."' (Los 2012)

The nominal subject typically follows the finite verb because it tends to be new information – this is why it is a full nominal rather than a pronoun. In (3), Ioseph, although not discourse-new, is no longer activated, as there has been an interpolation, and hence requires re-activation by the use of his name. Note the adverbial *On þam*, which constitutes a link with the preceding discourse:

(3) Þa dyde man hig on cwearterne.(…).

Then did they them in jail. (…).

On þam *wæs* eac ***iosep*** gebunden [cogenesiC:191]

In that was also Joseph bound

'Then they put them in jail. (…) Joseph was also in that jail'

The findings in van Kemenade, Miličev and Baayen (2008) and van Kemenade and Miličev (2012) show that subject placement in OE was information-structurally motivated, as it is not only pronouns that appear in

the higher subject position, but also full NPs that have specific anaphoric reference, which suggests that subject positioning is determined by discourse/information status (cf. Bech 2001). Van Kemenade and Westergaard (2012) show that same holds in early Middle English.

1.2. The changing role of the English subject

The loss of V2 in the fifteenth century in this view is more than a loss in the frequency of a particular word order: it spells the end of clause-initial adverbials as unmarked discourse links (see also Hinterhölzl & van Kemenade 2012; van Kemenade 2012). The canonical order of PDE sentences has the subject in the clause-initial slot, as the only information-structurally neutral way to start a clause (Downing & Locke 2002; following Halliday 1994 [1985]). PDE clause-initial adverbials do in fact occur, but they are less common (Biber et al. 1999: 802). They are also more restricted in their use than their OE counterparts, in that they tend to be forward-looking rather than anaphoric, and could perhaps be regarded as temporal or spatial frame-setters determining for which time and place the following proposition applies rather than links to the previous discourse, as in (4b), where they are also contrastive.

(4) a. How is business going for Daimler-Chrysler?

b. [In GERmany]$_{Frame}$ the prospects are [GOOD]$_{Focus}$,
but [in AMErica]$_{Frame}$ they are [losing MOney]$_{Focus}$. (Krifka 2007: 46)

As a result of the restrictions on the use of clause-initial adverbials, the PDE subject has acquired a greater functional load. This is illustrated by the PDE translations of the ME sentences in (5). The clause-initial PP *with this money* in ME (5a) would appear in PDE either as a subject (as in (5b)), in a cleft (as in (5c)), or as an object (as in (5d)). It is the subject that performs the task of discourse linking – or ensuring discourse cohesion – here, as it is the function-of-choice to encode given information. The change to SVO canonical word order in early Modern English (eModE) introduced the mapping of syntactic function with information status: subject with given information, object with less given or new information.[2]

The increasing restriction on first position adverbials is not only suggested by quantitative evidence for PDE as in Biber et al. (1999), but also by qualitative evidence.[3] Adverbials of time and place may easily be

[2] "New" information can also take the shape of a new *relation* between constituents that have already been introduced as mental entities in the discourse model (Lambrecht, 1994).
[3] Although Biber et al make no claim about the historical development, the small percentages of clause-initial adverbials he finds for PDE compared with the percentages for OE in the historical corpora does indicate such a development.

interpreted as frame-setters, and hence do not particularly stand out in PDE as different from their OE equivalents. Adverbials of means (or instrument) encoding discourse links are a different matter: they are less likely to be acceptable as frame-setters, and hence are more marked in first position in OE/ME/EModE/PDE comparisons. Compare the literal PDE translation of ME (5a) ('with this money, the pope renovated the Capitol') and the other ways which PDE has available to express the same idea – a subject in (5c) ('This money') and (5d) ('This') or an object in (5d).

(5) a. In þis tyme was founde [a gret summe of mony]$_i$ at Rome in a rotin wal (...). [With þis mony]$_i$ þe pope ded renewe þe Capitol and þe Castell Aungel. [cmcapchr:3763-8]

b. [This money]$_i$ was used by the pope to renew the Capitol and the Castel Sant'Angelo.

c. [This]$_i$ is [the money that was used by the pope to renew the Capitol and the Castel Sant'Angelo].

d. The pope used [this money]$_i$ to renew the Capitol and the Castel Sant'Angelo.

If discourse links like *with this money* in (5a) are increasingly expressed by means of a subject, rather than a clause-initial adverbial, we would expect much more switching between subjects, because subjects are no longer

reserved for protagonists. Note that such discourse links are often inanimate entities, as in (5), again in contrast with protagonist subjects. If the subject, rather than an adverbial, is increasingly used for linking, we expect to find an increase of inanimate subjects over the eModE period.

There is a second reason why inanimate subjects may be expected to increase over time. Psycholinguistic studies comparing PDE and ModG online retellings of video clips reveal a difference in the narrative perspective taken by speakers that may also be relevant to OE. These comparisons show that PDE uses the subject to not only encode protagonists, but also non-protagonist, inanimate forces, such as *the wind* in (6b) (Carroll & Lambert 2005; Carroll et al. 2004), whereas ModG speakers tell the story from the perspective of the protagonist, as is the case in (7a) and (7b). The PDE retellings may have non-protagonists as subjects, like *the wind* in (6b), while the ModG retellings keep the protagonist in subject position, often not mentioning *the wind* at all.

(6) a. A young man is surfing. (Carroll et al. 2004: 190)

b. The wind is blowing him off the board.

(7) a. Ein junger Mann surft auf hohen, schäumenden Wellen.

'A young man surfs on high, foaming waves.'

b. Dann wird er plötzlich vom Brett geweht.

'Then he is suddenly swept from the board.'

(Carroll et al. 2004: 190)

These comparisons also bring out another point about subjects: it seems plausible that the relative stability of the subject as a locus for the protagonist in ModG would result in a higher frequency (than in PDE) of clauses with the protagonist as subject, and hence to a higher degree of subject ellipsis in ModG (Carroll et al. 2008). As ellipsis, i.e. conjoined subject deletion, also occurs relatively frequently in OE (Fischer et al. 2000: 38-39), we hypothesize that it could be for the same reason: if the subject position is primarily reserved for protagonists rather than for discourse links as in (5b) or non-protagonist entities as in (6b), it is more likely to stay activated throughout long stretches of discourse and more easily recoverable when

ellipted. This observation is supported in the small pilot study of a comparison of an OE and a PDE retelling of the *Joseph in Egypt* story in Los (2009).

1.3. Hypotheses

The present study attempts to go beyond Los (2009) and test these hypotheses about the increased functional load of PDE subjects in a larger corpus. The discussion in the previous section may serve as a basis for four hypotheses:

(8) (i) *Ellipsis*

If the subject in OE is typically reserved for protagonists, it will be relatively stable and easily recoverable in ellipsis. The subject will become less stable as it becomes more functionally versatile, resulting in a *decrease in subject ellipsis (‘conjoined subject deletion’)* over time;

(ii) Referent switching

If the subject in PDE is no longer typically reserved for the protagonist, but also encodes non-protagonists (like *The wind* in (6b)) and discourse links (like *The money* in (5b)), there will be an *increase in subject referent switching* over time;

(iii) Subject animacy

If the subject in PDE is no longer typically reserved for the protagonist, but also encodes non-protagonists and discourse links, there will be a *decrease in the relative number of subjects referring to animate referents* over time.

(iv) Pre-subject linking

One of the forces contributing to an increase in subject functionality is the loss of coherence strategies available in the pre-subject position in OE, which manifests itself in a *decrease in pre-subject constituents having an unmarked link with the preceding discourse*.

2. Corpora

The approach we take to verify the hypotheses in (8) varies per hypothesis and is described in section 3. Our research is based on the collection of syntactically parsed corpora of historical English texts (see section 5 for a full listing of these corpora). These corpora provide us with information about the syntax of clauses and the parts of speech of clause elements. Some of the hypotheses in (8), however, can only be checked if coreferential information is available, i.e. information that gives us the referential status of each NP and a pointer to its antecedent, if there is one. This is why we have been enriching a growing subset of the texts available in the parsed corpora of English by providing them with coreferential information through the help of the program Cesax (Komen 2011).

This section briefly introduces the kind of referential information with which we have been enriching the existing texts, and then gives an overview of the enriched texts that are available at this moment.

2.1. Referential status

Speakers and hearers negotiate a Common Ground by each constructing a "mental model" of the situation presented in the discourse, a kind of mental stage, that is continually being updated (Garnham 2001; Johnson-Laird 1983; Zwaan & Radvansky 1998). Speakers and hearers keep track of the various referents that appear on the stage, and their attention is turned from one referent to another by various linguistic mechanisms: topic introducers, markers of foregrounding, backgrounding, and accessibility. These mechanisms are language-specific and hence also likely to change over time. Demonstratives in Old English, for example, not only mark definiteness in a more articulate way than an invariant definite determiner like *the* in PDE, they also constitute an alternative strategy of pronominal reference when used independently. Because we cannot rely on stable linguistic signs to signal referent tracking or Common Ground management through the various stages of English, we have chosen to research information structure by annotating corpora for referential information only, and then deriving *information structure* by combining syntactic and referential information. The referential annotation links every NP in a syntactically parsed corpus to an antecedent if it has one, and labels information about the nature of the link, i.e. the referential status: is it one of identity, like *Sue – she – his sister*, or is the link less direct, as in *the house – the kitchen*? (Komen 2012; Komen 2013). We distinguish five possible referential states, given in (9), which largely coincide

with the referential states used by the the PROIEL project (Haug et al. 2009).

[4]

(9) Referential state categories

a. Linked

i. Textual

1. Same entity → Identity

2. Different entity → Inferred

ii. Non-textual → Assumed

b. Unlinked

i. Non-referrable → Inert

ii. Referrable → New

We will refer to these five referential states as the *Pentaset*. These five categories are our "primitives", which, in combination with the syntactic information already present in the corpus, correlate with the traditional given-new distinction. The pentaset-annotation scheme allows large stretches of text to be annotated in relatively little time and is reliable as to interrater

[4] The PROIEL group uses five states: OLD, ACC-sit, ACC-inf, ACC-gen and NEW. These states largely coincide with the states in (9): "Identity" equals OLD, "New" equals NEW, "Inferred" equals ACC-inf, and the category "Assumed", which is discourse-new/hearer-old information, combines ACC-gen (general world knowledge) and ACC-sit (participants and props available in the extralinguistic context of the discourse, which includes deictic references such as *this story* in a sentence like "This story tells us how king Edmund died"). The state "Inert" does not have an equivalent in the PROIEL set of states.

agreement.[5] Although only a selection of texts have been annotated so far, the following sections will demonstrate how hypotheses such as those in (8i-iv) can be tested.

The text in (10) serves as an example to explain the Pentaset categories.[6]

[5] A comparison of referential state and antecedent annotation between three of the authors yields Cohen's kappa values between 0.84 and 0.88.

[6] The textual examples in this paper are taken from the parsed English corpora (see section 5) and referred to by their filename followed by the line number they occur in.

(10) a. [$_{NP}$ **I**] am the second son of [$_{NP}$ **a family of eight**], - six sons and two daughters, -

b. and was born on December 6, 1824, at [$_{NP}$ **Plymouth**], where [$_{NP}$ **my**] father and mother were on a visit after one of [$_{NP}$ **his voyages to India**].

c. My father was one of three sons of Captain J. Fayrer:

d. [$_{NP}$ **the eldest**] was the Rev. Joseph Fayrer, rector of St Teath, Cornwall;

e. the third, Edward, a midshipman in [$_{NP}$ **the navy**], was drowned when H. M. S. Defence foundered, with all hands, in a gale of [$_{NP}$ **wind**] in the Baltic in 1811.

f. My mother was the only daughter of a Lancashire gentleman named Wilkinson:

g she was descended on the female side from John Copeland, who took David, King of Scots, prisoner at [$_{NP}$ **the battle of Neville's Cross**]. [fayrer-1900:7-13]

The first constituent *I* in (10a) is discourse-new but addressee-old information, which receives the category of "Assumed" in the Pentaset; this kind of information leads to the creation of a mental entity in the mental model, linking it to the available extra-textual antecedent. Other constituents

with the same category are, for instance, *Plymouth* in (10b), *the navy* in (10e) and *the battle of Neville's Cross* in (10g).

The status of *a family of eight* is not only new to the discourse, but also to the addressee, for which reason it receives the category of "New" in the Pentaset; it leads to the creation of a new mental entity in the model, which is built up dynamically.

The personal pronoun *my* has an antecedent in the discourse (the pronoun *I* in the first line), and the entity referred to by the current constituent and its antecedent completely coincide, so that they receive the Pentaset category of "Identity".

The constituent *the eldest* in line (10d) refers back to *three sons of Captain J. Fayrer* in (10c), but the entities are not identical – they stand in a part-whole relationship. This relationship as well as other bridging inferences receive the Pentaset category of "Inferred" (see e.g. Irmer 2011; Prince 1981).

There is one final Pentaset category called "Inert", and the noun phrase *wind* in (10e) is an example of it. This *wind* really is an attribute to *gale*, so that, as attribute, it cannot refer to something, nor can it be referred to. In other words: such noun phrases are inert to the whole process of referencing; no separate mental entity is created for them.

The information status of a noun phrase like *his voyages to India* in (10b) would be "New" as far as the Pentaset is concerned, since the information is both new to the addressee as well as to the discourse, and a

new mental entity needs to be set up in the mental model. A finer-grained system, such as Prince's (1981) the "taxonomy of given and new", would assign it the "Brand-new anchored" status. However, this finer-grained distinction is derivable from the available syntactic information and the Pentaset statuses, which is an important point we would like to stress: it is the combination of syntax and referential states that lead to information status. The status of "Brand-new anchored" can be assigned to any constituent that (a) has the Pentaset category of "New", and (b) contains at least one constituent with the Pentaset status of "Identity". In the current example the pronoun *his* has the status of "Identity", since its antecedent is *my father*, and the entity referred to by *his* and *my father* is identical.

It should be noted that coreferential chains consist of only those references to a participant that can be linked together with the category "Identity".

2.2. Enriched texts

The "Cesax" program (Komen 2011; 2012) has been instrumental in semi-automatically adding referential status features to each NP and, where applicable, providing a pointer to the NP's antecedent. The texts that have until now been enriched with coreferential information are listed in **Table 1**.

Table 1. Texts that have been enriched with coreference information[7]

File	Period	Word count	Genre
coapollo.o3	OE: O3	6545	Fiction
covinceB	OE: O14	728	Biography
Coeuphr	OE: O14	3658	Biography (saint's life)
cmsawles.m1	ME: M1	4111	Homily
cmkentse.m2	ME: M2	3534	Homily
cmhorses.m3	ME: M3	8902	Handbook
cmreynar.m4	ME: M4	8850	Fiction
fisher-e1-h	eModE: E1	4853	Sermon
fabyan-e1-h	eModE: E1	5478	History
perrot-e2-h	eModE: E2	4831	Biography
behn-e3-p1	eModE: E3	5908	Fiction
jpinney-e3-p1	eModE: E3	186	Letter
brightland-1711	lModE: B1	1341	Educ_Treatise
defoe-1719	lModE: B1	9378	Fiction
fleming-1886	lModE: B3	9038	Handbook
long-1866	lModE: B3	8851	History
skeavington-184x	lModE: B3	9132	Handbook

The texts that have been enriched come from different subperiods of the four main periods of the English language: three texts from the OE period, three

[7] The period abbreviations used in this article are: OE (450-1150), O1 (450-850), O2 (850-950), O3 (950-1050), O4 (1050-1150), ME (1150-1500), M1 (1150-1250), M2 (1250-1350), M3 (1350-1420), M4 (1420-1500), eModE (1500-1710), E1 (1500-1569), E2 (1570-1639), E3 (1640-1710), lModE (1700-1914), B1 (1700-1769), B2 (1770-1839), B3 (1840-1914). The sub period "O14" means that the OE manuscript is from the 4th (final) subperiod of OE, but the original text could have been from any time within OE, starting with O1.

texts from the ME period, four texts from the eModE period and five from the late Modern English (lModE) period. Although these texts do not all belong to the same genre, we selected them on the basis of their narrative style: all texts have one or multiple clear protagonists and make up a single narrative – although some of them are divided into chapters.

3. Experiments

3.1. Subject ellipsis

Subjects that are ellipted under conjunctions are easily recognizable in the parsed corpora of English (even without additional referential information): the subject NP carries the "normal" subject label (the label is NP-NOM for Old English and NP-SBJ for the other English periods), but it is also an endnode with a text value marked as *con*. The algorithm we use in order to search for subject ellipsis in the corpora is described in (11).

(11) Subject ellipsis algorithm

Step 1: Consider each NP in the text, and check if it satisfies the conditions:

Condition a: the NP label is the label for a subject

Condition b: there is only one daughter, and this daughter is the text *con*

Step 2: Check if the NP is the daughter of a main clause or subclause

The first step in the algorithm checks whether the NP has the correct value, as explained above, while the second step checks to see if the NP that is found is actually part of a finite clause: a main clause (where the parent of the NP should have the label IP-MAT) or a subclause (with an IP-SUB parent).

The query to find instances of ellipted subjects has been run on all four parsed corpora of English described in section 2. The number of ellipted subjects thus found has been compared with the number of sentences in main clauses and subclauses actually containing a "proper" subject.[8] We define

[8] We have excluded non-overt subjects that are marked as *traces* in the parsed English corpora. This kind of subject frequently occurs in relative clauses, such as (i):
(i) I will however be thankful for the blessings$_i$ [$_{IP\text{-}SUB}$ that t_i are spared to me].
[reeve-1777:48]
The subject of the relative clause is the trace t_i, which links to the antecedent *the blessings*. Clauses such as these are not examples of ellipsis in the usual sense.

proper subjects are those that are either lexically realized on the surface or ellipted. The results are shown in Figure 1.[9]

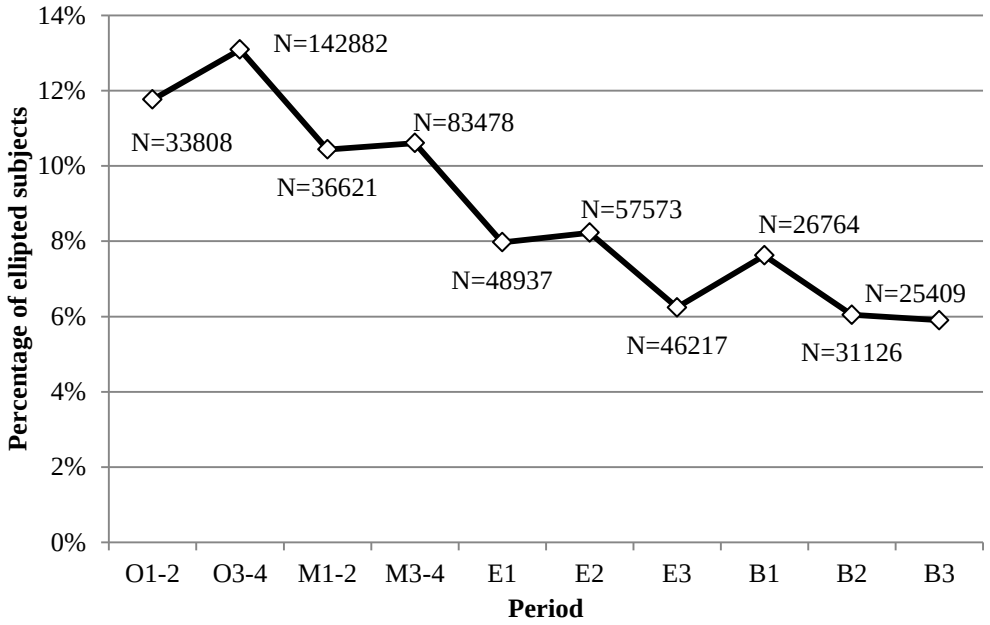

Figure 1. Ellipted subjects

Figure 1 shows a clear decline in subject ellipsis between OE and lModE, which is in line with our hypothesis as described in (8i). The question that prompts itself here is how the distribution of NPs changes in general. To that end we have conducted experiments, again on all of the four parsed corpora of English, where we have looked at the division of the NPs according to their type. We have divided the NP types of the subjects into four categories: (a) pronominal subjects, (b) subjects that are ellipted under conjunction, (c) expletive subjects, and (d) all other (lexical) NP subjects. Figure 2 shows the

[9] The transitions are significant according to the two-tailed Fisher's exact test (p<0,01), except for: M1-2 to M3-4 ($p=0.38$), E1-E2 ($p=0.13$) and B2-B3 ($p=0.47$).

distribution of the subject types when the subjects have been restricted to those that are subjects of finite clauses (main clauses and subclauses).

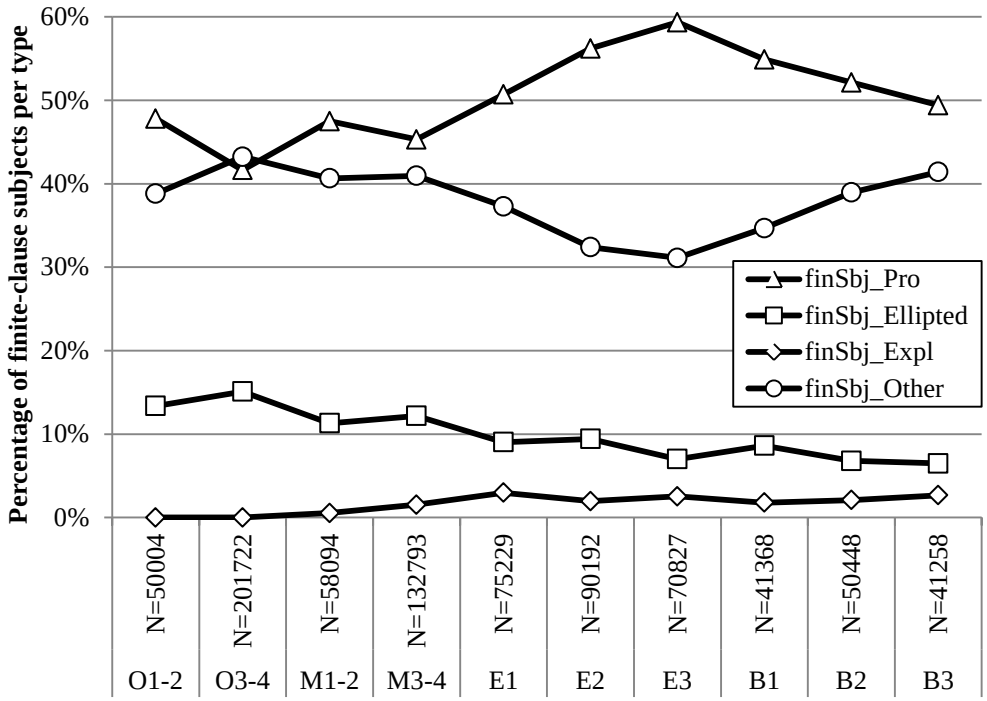

Figure 2. Finite clause subject type distribution

What we see in Figure 2 is, first of all, the ellipted subjects from Figure 1, but now on a more compressed scale. We also see a slight increase in expletive subjects, particularly over the ME period.[10] The relative number of pronominal subjects vacillates over the different periods, but eventually the lModE percentage is only slightly higher than the OE percentage.

[10] All the transitions are significant according to the two-tailed Fisher's exact test (p<0,05) except for these:
Pro - (all transitions significant)
Ellipted - the transitions to M3-4 ($p=0.38$), to E2 ($p=0.13$) and to B3 ($p=0.47$).
Expl - the transition to O3-4 (p=1,00)
Other - the transition to M3-4 ($p=0.08$)

We conducted a follow-up experiment in which we not only restricted ourselves to finite clauses, but also stipulated that the subjects had to be more "protagonist-like". We defined "protagonist" as third-person non-neuter discourse participants about whom information is given in the text.[11] For this reason, we excluded all first-person and second-person subjects – which might refer to the narrator and the reader of the text respectively – as well as the third-person neuter singular subjects. The results of this experiment, again for all the syntactically parsed corpora, are shown in Figure 3.

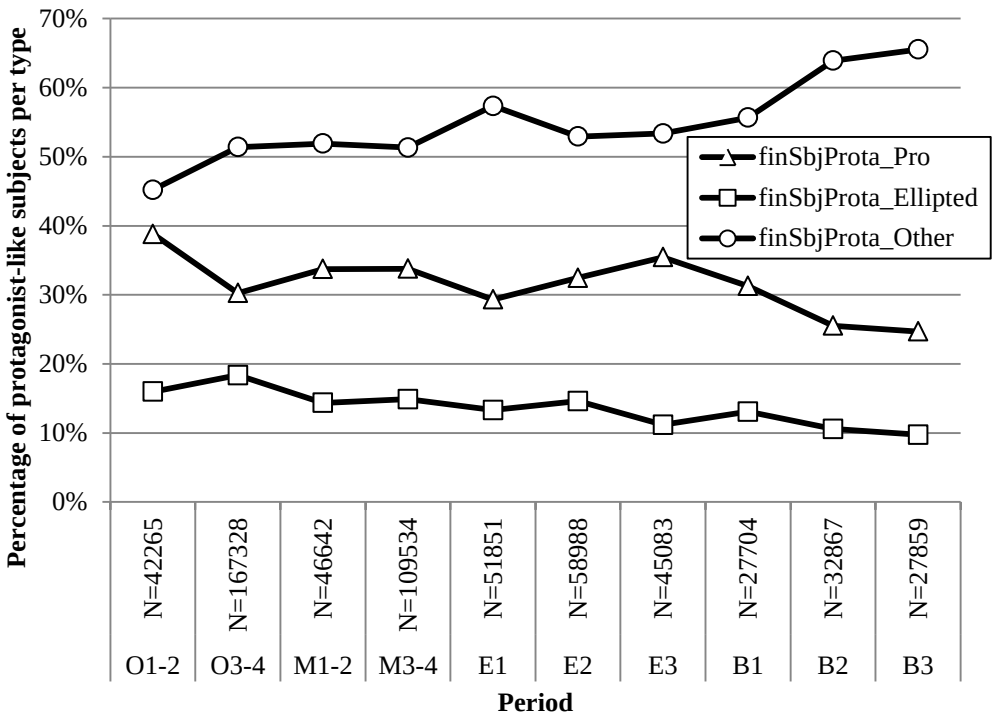

Figure 3. Distribution of finite-clause protagonist-like subjects according to their NP type

[11] The "non-neuter" stipulation does not work for OE, which has grammatical gender.

Our exclusion of third-person neuter singular subjects leads to the loss of expletives in Figure 3. The ellipted subjects are still visible, but their decrease is slightly less pronounced. We also see that there is a decrease from 38% to 25% in protagonist-like subject pronouns (and, conversely, the more lexical NPs, labelled as "Other", increase from 45% to 65%).[12]

This development is completely in line with our hypothesis in (8i). As the functional load of the subject increases (encoding more discourse links and more non-protagonists), the proportion of subjects encoding protagonists decreases. When, as a consequence of the increased number of functions the subject has to fulfill, the subject-referent switches more often (between discourse links, non-protagonists and protagonists), we may need to "reactivate" the referent more frequently by using a nominal NP (e.g. a proper name – *Sue* – or a definite article plus a noun – *the woman* – or a possessive pronoun plus a noun – *his sister*) instead of a pronoun. This means that we would expect the ratio of pronouns/nominal NPs to encode protagonists to decrease.

3.2. Subject referent switch

[12] All the transitions are significant according to the two-tailed Fisher's exact test ($p<0,05$) except for these:
Pro - transition to M3-4 ($p=0.91$)
Ellipted - transition to M3-4 ($p=1,00$) and to B3 ($p=0.08$)
Other - transition to M1-2 ($p=0.08$), to M3-4 ($p=0.92$) and to E3 ($p=1,00$)

3.2.1. A definition of subject-referent switch

Subject-referent switch occurs when the subject referent of one clause differs from that of the previous clause. An example of subject-referent switch is in (12), where the subject changes from *John* in (12a) and (12b) to *his daughter* in (12c) and (12d).

(12) a. [$_{sbj}$ John$_i$] entered the room where [his$_i$ daughter$_j$] usually watched television.

b. [$_{sbj}$ He$_i$] looked around and [$_{sbj}$ 0$_i$] saw his$_i$ daughter$_j$, [$_{sbj}$ who$_j$] was sitting on the couch.

c. [$_{sbj}$ She$_j$] looked up and [$_{sbj}$ 0$_j$] made a face at him$_i$ as [$_{sbj}$ he$_i$] passed by.

d. [$_{sbj}$ She$_j$] had had a rough day at school.

Quantification of subject-referent switching for a whole text can be obtained by comparing the number of subject-referent switches that occur with the total number of subjects, as in formula (13). This formula compares the number of subject-referent switches occurring with "n-1" – the number of sentences minus one – for the simple reason that subject-referent switching cannot be measured for a text consisting of just one sentence. It is for this same reason that the subject number *i* starts with sentence number two.

(13) Subject-referent switch definition

$$SRS = \frac{\sum_{i=2}^{n} Ref_{Sbj_i} - Ref_{Sbj_{i-1}}}{(n-1)}$$

While this general formula suffices to quantify the relative number of subject-referent switches occurring in a text, there are two restrictions we adhere to: one for the subject and one for the kind of sentences we count. These restrictions are given in (14).

(14) a. **Subject:** Include all explicit subjects as well as subjects ellipted under coordination.

b. **Sentence type:** Include main clauses (marked as IP-MAT) as well as subordinate clauses (marked as IP-SUB), but do not include relative clauses.

The first restriction has to do with **subjecthood**. Which subjects are relevant to our hypothesis? We should at least accept all *explicit* subjects, i.e. all subjects that are expressed overtly. But should we also include ellipted subjects, such as the "0" subject in (12b) referring to *John*, and the "0" subject in (12c) referring to *his daughter*? As an ellipted subject necessarily refers to the same participant as the subject in the preceding sentence, subject-referent switch cannot occur with an ellipted subject. This means that ellipted subjects

are strictly speaking not relevant to the hypothesis. However, excluding ellipted subjects without excluding the sentences containing ellipted subjects would skew the data. We have therefore opted to include ellipted subjects, as they are available in the syntactically annotated corpora, and as they are included in the coreferential chains created by Cesax.

The second restriction has to do with the notion of **sentence**. For instance, (12b) could be considered one single sentence. However, it contains three clauses – each with its own subject: *he*, *"0"*, and *who*. These clauses are the result of coordination and subordination. The question is whether all coordinated and subordinated clauses should be included in this study. Relative clauses are embedded in a main-clause NP, and in the majority of cases pertain to the referent of their antecedent, making them 'dead ends' in a chain. However, this is not the case for some non-restrictive relative clauses. As these non-restrictive clauses cannot be filtered out of the group of relative clauses as a whole, we decided not to include any relative clauses in this study. Coordinated clauses and other subordinated clauses are included.

3.2.2. Measuring subject-referent switch

All expressions in a narrative that refer to one particular participant together make up a coreferential chain; each instance in the chain has exactly the same identity. **Table 2** visualizes the coreferential chains for participants "John" and "his daughter" from the narration in (12). If we look at the coreferential

chains for individual participants, the change in grammatical role – a change from subject to some other role – on one chain does not necessarily tell us anything about the subject-referent switch that takes place between two clauses. The change in John's grammatical role from *Subject* to *PossDet* in line (12b), for instance, is not related to the subject-referent switch taking place between *John* in (12b) and *his daughter* in (12c).[13]

Table 2. Coreferential chains of the participants in (12)

Line	**Clause**	**John**		**His daughter**	
		Form	**Role**	**Form**	**Role**
D	Main	-	-	She	Sbj
c3	Sub	he	Sbj	-	-
c2	Main	him	PPobj	0	Sbj
c1	Main	-	-	She	Sbj
b3	*RC*	-	-	*Who*	*Sbj*
b2	Main	his 0	PossDet Sbj	daughter	Obj
b1	Main	he	Sbj	-	-
a2	*RC*	*his*	*PossDet*	*daughter*	*Sbj*
a1	Main	John	Sbj	-	-

[13] One reviewer wondered whether we had included mentions of participants in direct speech. We have, as there is no reason to assume that including those instances would skew the data.

What is needed for the proper calculation of subject-referent switching is an algorithm that walks every allowable clausal domain (main clauses and subordinate clauses, excluding relative clauses), and calculates the number of times the referent of the subject changes. Such an algorithm needs to recognize which referent each subject in subsequent clauses refers to. This information can be derived from the syntactically annotated corpora that have been enriched with coreferential information, since each NP receives as a feature a numerical ChainId that uniquely identifies the chain it belongs to. The algorithm that calculates subject-referent switch is described in (15).

(15) Subject-referent switch algorithm

Step 1: Consider each NP in the text, and check if it satisfies the conditions:

Condition a: the NP label is the label for a subject

Condition b: the NP is not a "Trace"[14]

Step 2: Check if the NP is the daughter of a main clause or subclause (not a relative clause)

Step 3: Let `$chid` be the ChainId value of this NP

Step 4: If `$chid` is not equal to `$lastid`, then output this instance

Step 5: Let `$lastid` be the current `$chid`

This algorithm considers all the NPs that can be found in the text one by one, and checks whether a given NP conforms to two necessary conditions: (a) it is a subject, and (b) it is not a trace. Step 2 checks whether the NP is the daughter of a main clause or a subclause, excluding relative clauses. Once we are satisfied with the basic conditions, we can go through steps 3-5 to see whether a switch in chain has taken place (the value of ChainId then differs from the last value we have stored). If this is so, we put the NP in the output. Once the algorithm has done its work, we can count all the NPs in the output,

[14] Since relative clauses are excluded (see the end of section 3.2.1), no relative clause traces will be encountered. The parsed corpora do, however, contain other traces (e.g. *wh*-clauses, A and A' movement), and we exclude all of these categories in the current algorithm.

divide this by the number of main clauses and subclauses (with non-trace subjects), and we end up with the average subject-referent switch.

3.2.3. Subject-referent switch results

The subject-referent switch algorithm described in (15) has been run on the enriched text corpus. The results are shown in Figure 4.

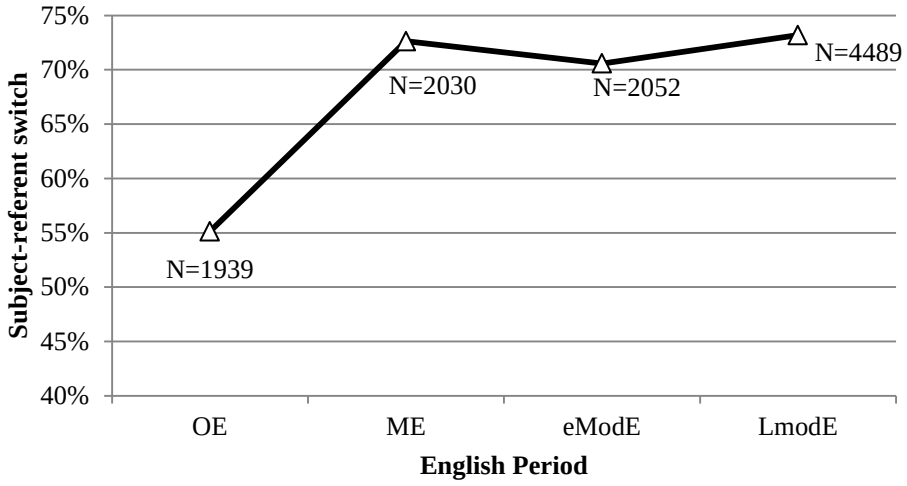

Figure 4. Relative number of main clauses and subordinate clauses featuring subject switch

The numbers (see **Table 3**) show an increase between OE to lModE, although there appears to be either a peak in ME or a dip in eModE.

Table 3. Relative number of main clauses and subordinate clauses featuring subject switch[15]

	OE	**ME**	**eModE**	**lModE**
Clauses	1250	1176	1203	2592
Subject switches between clauses	689	854	849	1897
	55.1%	72.6%	70.6%	73.2%

The general increase in subject-referent switching between OE and lModE is in line with hypothesis (8ii). The rise in ME which is then followed by a slight fall in eModE remains unexplained. The total number of clauses available for subject switch to happen differs between periods, which might influence the significance of the results. More data from OE, ME and eModE in particular is needed. Another cause of the unexpected trend witnessed here may be sought in the genre differences between the texts (see **Table 1**). Some of the texts are narrated from a first-person singular perspective, which could logically be a trigger of subject-referent switch. Such patterns may become more obvious when more texts have been annotated than at present. What this section shows is the kind of information that can be extracted from a referentially and syntactically annotated corpus.

[15] The transition from OE to ME is significant according to the two-tailed Fisher's exact test (p<0,05), while the other transitions are not; the transition to eModE has $p=0.66$, and the one to lModE has $p=0.52$.

3.2.4. Subject chain distribution

One question that comes to mind when we look at subject-referent switching is whether the length-distribution of chains that contain a subject changes over time as a result of the increase in subject switching. Is it just that we have fewer long chains in lModE? Or are we getting more short chains? Or both? In order to answer these questions, we have conducted an experiment on the enriched corpus, where we note the distribution of those chains that contain at least one constituent functioning as a subject in a finite clause. The algorithm runs along the lines in (16), yielding the results shown in Figure 5.

(16) Subject chain distribution algorithm

Step 1: Consider each NP in the text, and check if it satisfies the conditions:

Condition a: the referential type is such that this starts a chain (the Pentaset status is “Assumed”, “Inferred” or “New”)

Step 2: Check if there is an NP on the chain started in step 1 that satisfies:

Condition a: the NP is not a trace

Condition b: the NP is a subject

Condition c: the NP is the daughter of a main clause or subclause

Step 3: Store the length of this chain

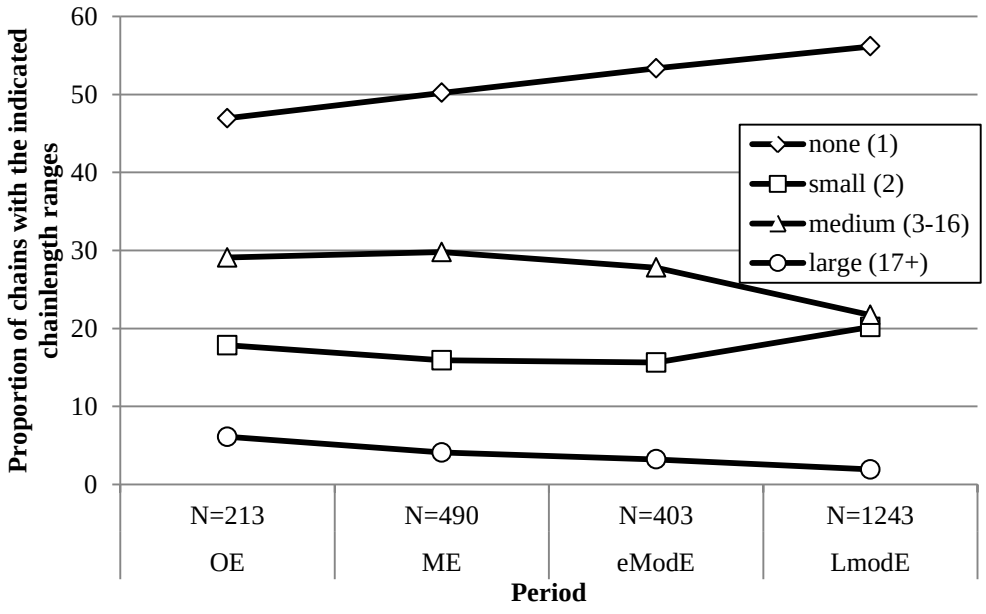

Figure 5. Distribution of chains having at least one subject[16]

What we see is that there are indeed changes in the distribution of the chain lengths. The relative number of larger chains (those with 17 or more constituents) decreases steadily from OE (6%) into lModE (2%). The slightly smaller chains, with lengths from 3-16 constituents, also decrease. Their contribution is 29% in OE and only 22% in lModE. The relative contribution of the "non-chains" (which are constituents that are not referred to at all, indicated by "none" in the picture) increases steadily from 47% in OE to 56% in lModE. The downward trends of the longer chains and the upward trends of the smaller chains both support the picture we have been sketching, in which the increasing functional load of the subject leads to an increase in the number of subject-referent switches, partly because of an increase in the number of short-lived subjects (those that have no chain, or only one element on the chain).

3.3. Subject animacy

The hypothesis on subject animacy in (8iv) states that we expect an increase in inanimate subjects over time. In order to measure this, we need to do some

[16] The *p*-values of the transitions according to the two-tailed Fisher's exact test are as follows:
None(1) - transition to ME (*p=0.*46), to eModE (*p=0.*38), to lModE (*p=0.*33)
Small(2) - transition to ME (*p=0.*58), to eModE (*p=0.*93), to lModE (*p=0.*05)
Medium(3-16) - transition to ME (*p=0.*93), to eModE (*p=0.*55), to lModE (*p=0.*01)
Large(17+) - transition to ME (*p=0.*25), to eModE (*p=0.*59), to lModE (*p=0.*17)

additional enrichment. The parsed corpora of English contain word and phrase level syntactic categories, but no animacy information. The texts we enrich with the Cesax program get referential information, but they do not have animacy added either. However, in the process of deriving the referential information with Cesax, the NPs in the texts are also enriched with a PGN feature that gives their person, number and (grammatical) gender.

3.3.1. Determining subject animacy

In order to verify the subject animacy hypothesis, we have opted to semi-automatically add animacy information to two texts – one from OE and one from lModE. The semi-automatic process first attempts to determine animacy based on the available syntactic and PGN information. If it fails to get a result, it will ask the user to choose between “animate”, “inanimate” and “unknown”. The animacy determination process works on texts that have already been enriched with referential information, and follows the algorithm in (17).

(17) Animacy determination

Step 1: For each NP *x* in the text that has no animacy yet

Step 2: For each NP *y* on the chain of *x*

Step 3: Try to get the animacy of *y*:

Situation a: (not OE) PGN is first or second person

or 3fs or 3ms[17] → animate

Situation b: (not OE) PGN is 3ns → inanimate

Situation c: NP is vocative → animate

Situation d: NP is a temporal, measure, number

or nominalized clause → inanimate

Situation e: head-noun has known animacy →

copy animacy of head noun

Situation f: head-noun ends on nominalization suffix →

inanimate

Step 4: If animacy unknown →

ask user for animacy of last *y* constituent

Step 5: Spread the animacy of *y* to all constituents on the chain of *x*

[17] Since OE has grammatical gender, we checked each 3fs, 3ms and 3ns referent in order to determine whether the referent was animate or not.

The algorithm methodically addresses each NP in the text (step 1), and when it finds an NP that has no animacy assigned yet, it tries to determine the animacy of the whole chain of which this NP is part by getting the animacy of one constituent on the chain (step 3). The person information (first and second person versus third person) gives some indication of animacy, as does the gender information, if available. Situations c-d in (17) show that syntactic information can sometimes help in deriving animacy.[18] Situations e-f in (17) deal with the head noun of the NP. If this head noun has already been encountered elsewhere in the text, the animacy can simply be copied, and if not, there are still some clues in the form of the head noun, such as the presence of a nominalization suffix (e.g. *–ion*, *-ity* etc). If all these measures fail, the algorithm asks the user to make a decision (step 4). The final part of the algorithm spreads the result to all the elements on the chain of which the NP we started out with is part, since all the elements on an (identity) chain refer to the same participant or object, and therefore must have the same animacy feature.

3.3.2. Subject animacy results

We used the semi-automatic algorithm in (17) to add animacy features to one OE text (Apollonius of Tyre, coapollo) and one lModE text (Defoe, defoe-1719).

[18] What we refer to as "syntactic" information here is the information that can be gleaned from the syntactically parsed corpora of English. This not only includes word category (e.g. verb, noun), and phrase category (e.g. AdjP, NP), but often also functional information, such as NP role (subject, predicate, temporal, measure, vocative etc.), type of clause, etc.

Since the hypothesis in (8iv) states that we expect to see an increase in the percentage of inanimate participants in a text that are referred to in a subject position, we used a corpus research project described in the algorithm in (18) to determine (a) the total number of participants in a text that appear at least once as a subject, and (b) the number of these that is inanimate.

(18) *Inanimate subject algorithm*

Step 1: Consider each NP in the text, and check if the reference type is "New", "Inferred" or "Assumed"

Step 2: Check if the chain started by this NP has one constituent as subject

Step 3: Check the animacy of the NP

We start in step 1 by addressing each participant, by checking all NPs that can function as the start of a coreferential chain (the texts must be annotated in such a way that each participant is only part of one coreferential chain). Such NPs are characterized by having one of the three reference types stated in condition *a* (a "New" NP points to an entity that has not been mentioned before, but can potentially be referred to later, and "Inferred" NP relates to an already mentioned entity, but is not exactly the same, and it too can be referred to again, and an "Assumed" NP is an entity that is new in the text but assumed to be known to the addressee, and it too can be referred to again).

The NP we have as well as the constituents on the chain formed by the NP are checked until one of these is found that has the function of a clausal subject. This part of the algorithm gives us the base number: all participants that function at least once as subject in the text. The last step of the algorithm, step 3, checks the animacy of any NP that fulfills the preceding conditions (it is an NP on a chain that has at least one constituent as subject). It is here that we count all the inanimate participants in the text.

The corpus research project, which is the CorpusStudio (see Komen 2012) implementation of the procedure in (18), is executed on the OE and lModE texts mentioned above, yielding the results in **Table 4**.

Table 4. Animacy compared between OE and lModE[19]

Period and text	**Chains**	**Chains with subject**	**Inanimate ones**	**Inanimacy**
OE (coapollo)	848	126	55	43,7%
lModE (defoe-1719 + brightland-1711)	1356	307	168	54,7%

What we see here is an increase in the relative number of inanimate participants that function as subject at least once in a text. These results confirm the hypothesis in (8iv), but we must note that the sample size is very

[19] The significance according to the two-tailed Fisher exact test yields a value of $p=0.0348$ when we compare the inanimate chains with those chains that have a subject.

small. Future work on more annotated texts should help us get a clearer picture of the rise of inanimate subjects in English.

3.4. Pre-subject linking

One stimulus for the increased functionality of the subject is the loss of pre-subject constituents to function as unmarked discourse links, as argued in section 1.2. Clause-initial PPs or NPs in environments of the XP-S-V_{fin} serve less frequently as links to the immediately preceding context over the course of time (see also Hinterhölzl & van Kemenade 2012; van Kemenade 2012). This section describes an experiment where we measure this phenomenon by looking at the referential status and the antecedent distance of PPs and object NPs in the XP-S-V_{fin} environment. It is only because the texts we look at have been enriched with referential information (using Cesax) that we are able to quantify the changes.

3.4.1. Clause-initial linking

Clause-initial linking is a way of establishing paragraph-internal cohesion. The process of clause-initial linking has changed dramatically over time, in particular after the decline of the English demonstrative paradigm. Los and Dreschler (2012) looked at main clauses starting with a PP, which includes a wide range of environments (such as PP-S-V, PP-V-S, or more generally: PP-X). They manually investigated texts from OE to lModE, and found that the

proportion of clause-initial PPs containing a demonstrative pronoun drops from 17% in OE, to 4% in lModE.[20]

Another corpus study was performed on all syntactically parsed texts from OE to lModE, searching for main clause-initial constituents containing demonstratives or pronominal adverbs.[21] This study (see **Table 5**) shows a steady decline in the proportion of clause-initial constituents containing a linking element.

Table 5. Main clause-initial constituents containing a demonstrative or pronominal adverb

	English			
	OE	**ME**	**eModE**	**lModE**
matFirstConst	66425	56805	63969	39677
matFirstConst (Dword)	14441	8495	5443	2945
matFirstConst (Dadv)	12551	6278	4247	917
matFirstConst (Dadv + Dword)	**40.6%**	**26.0%**	**15.1%**	**9.7%**

Studies like these only take the grammatical category of the first contituent into account, and fail to involve its referential status. It is for this reason that

[20] This study only takes independent demonstrative pronouns into account (such as *that* in a PP such as *by that*), excluding demonstratives that function as determiner in a PP's NP object (such as *this* in *in this way*).
[21] Unlike the previously mentioned study, this study does include dependent demonstratives—those that combine with an NP. Pronominal adverbs are combinations of a demonstrative and a pronoun such as *therefore*, *thereby*, *therewith*.

the following section describes an experiment that does take the referential status of clause-initial constituents into account. It should be noted that the number of texts available for this type of research is limited, as only a small proportion of the syntactically annotated corpora has been enriched with coreferential information.

3.4.2. Determining pre-subject linking

The question whether PPs or argument NPs preceding the subject in XP-S-V_{fin} environments contain a link to the preceding context can be investigated in the enriched texts by looking at the referential status of the clause-initial XP. The two types of XPs require a slightly different treatment. We only want to look at those clause-initial NPs that are marked as direct or indirect objects (this is visible from the syntactic labels of the constituents). As for clause-initial PPs: we only want to look at PPs that have an overt NP object adjacent to the P. The pre-subject linking algorithm that takes these requirements into account is provided in (19).

(19) Pre-subject linking

Step 1: Consider each main clause in the text, and check if it satisfies the conditions:

Condition a: there is a clause-initial XP (a PP or argument NP)

Condition b: there is an overt subject NP

Condition c: there is a finite verb

Condition d: word order is XP-Subject-FiniteVerb

Step 2: Let *x* be the NP part of the clause-initial XP

Step 3: Determine the linking status of *x* as follows:

"Linking": the referential status of *x* is *Inferred* or *Identity* and the link is anaphoric

"None": the referential status of *x* is something else, or the link is cataphoric

The algorithm starts in step #1 by looking for main clauses. These main clauses need to have a clause initial constituent that is either a PP or an argument NP, as per condition #1a. The other conditions #1b and #1c state that a subject and finite verb also need to be explicitly present. The last condition #1d requires these elements to be present in the correct order.[22] Step #2 of the algorithm makes sure we continue to work with an NP – this is the

[22] The Xquery implementation of the algorithm requires immediate adjacency but excludes conjunctions and extralinguistic nodes like those marked as "CODE".

object NP if that happens to be the clause-initial XP or else it is the NP part of the PP. The last step #3 determines the status of the NP that has been identified in the previous step. The status of "Linking" is only assigned to those NPs that have a referential status of "Inferred" or "Identity" where the antecedent being referred is from the *preceding* context.

3.4.3. Pre-subject linking results

We have used an Xquery implementation of the algorithm in (19) in the CorpusStudio program (see Komen 2012) to look for the XP-S-V_{fin} environments in the enriched texts shown in **Table 1**. **Figure 6** shows how the proportion of clause-initial XPs with a link to the preceding discourse changes over time.

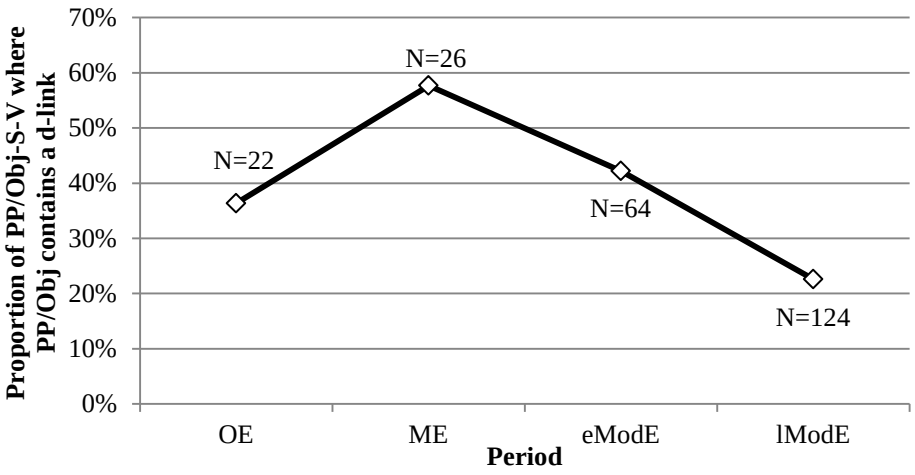

Figure 6. Decline of pre-subject constituents with a link to the preceding context[23]

Even though the number of XP-S-V_{fin} environments found for OE and ME is not very large, there is a clear trend that confirms our hypothesis. There is a steady decline from pre-subject linking elements from almost 60% in ME to a mere 22% in lModE, and the increase from OE to ME is not significant due to the small amount of data available for these periods.

(20) a. (Þe þridde is bounte þat is best of alle.)

And **þat** þou schaltknowe by þese signus. [cmhorses:23-25]

and that you shall know by these signs

[23] The *p*-values of the transitions according to the two-tailed Fisher's exact test are as follows: transition to ME (*p=0.*16), to eModE (*p=0.*24), to lModE (*p=0.*01).

'(The third is the character [of the horse], and this is the most important of all.)

*And you will know **this** by the following signs.'*

b. (þt heaued þrof is þe feont. þe meistreð ham alle.)

aġeines **him** & **his** **keis**. þe husebonde þt is

against him & his henchmen the husband that is

wit; warneð hishus þus. [cmsawles:26-27]

Wit guards his house thus

'(Their chief is the devil, who commands them all.)

*Against **him and his henchmen**, the husband, that is Wit,*

guards his house like this:'

c. (I got no Body to come back with me but the Supra-Cargo and two Men.)

and with **these** I walk'd back to the Boats. [defoe-1719:482-483]

The examples illustrate that pre-subject objects can provide an unmarked (i.e. non-contrastive) link to the preceding context, as in (20a), as can pre-subject prepositional phrases, as in (20b). This option seems to still be available in lModE, witness the example in (20c), but is use is receding, witness the numbers in **Figure 6**. The changes we see from ME to lModE onwards must have continued, given the fact that (20c) would no longer be wholly felicitous in PDE. The reason why OE deviates from this trend of decline in pre-subject

constituents with a link to the preceding context may have to do with the fact that the amount of data for the OE period is relatively small.

What we may conclude, then, is that the pre-subject XPs in the XP-S-V_{fin} environment are increasingly unlikely to encode an unmarked link to the preceding context. This loss in functionality to express paragraph-internal cohesion must have resulted in an increasing pressure on the grammar at large, and as we claim, the subject in particular, to come up with alternative strategies.[24]

4. Conclusions and discussion

We hypothesized that the verb-second constraint in Old English and Middle English made a clause-initial position available that was multifunctional, both syntactically and information-structurally. Its many functions included providing a link to the previous discourse. The loss of V2 in the fifteenth century appears to have affected the status of first-position adverbials, which no longer could encode discourse links. Our hypothesis is that the subject took over some of the discourse linking functionality that was lost. Searching parsed corpora that have been further enriched with referential information allowed us to test this hypothesis by four experiments. The first experiment

[24] We are not claiming that it is always *grammar* that has to come with strategies for pragmatic notions such as cohesion; it is the language as a whole that will seek compensating strategies. Some of these may simply be lexical ones.

looked at conjoined subject deletion, and confirmed our hypothesis: as the functional load of the subject increases (encoding more discourse links and more non-protagonists), the proportion of subjects encoding protagonists decreases, as is visible in a decrease in conjoined subjects.

Our second experiment tested the hypothesis that, with the subject no longer "reserved" for the protagonist but also encoding discourse links or non-protagonists, we would expect the number of subject-referent *switches* to increase. This hypothesis was borne out, particularly for the transition from OE to ME. The proportion of subject-referent switches remains stable from ME onwards, which is contrary to what we might expect if we assume a direct relationship between subject-referent switching and the loss of V2. However, looking at the *referent chains* containing a subject, we find that the proportion of zero-length chains increases substantially from ME onwards, which is what we would expect if one of the functions of the subject increasingly becomes that of encoding one-time referents.

In the third experiment, we tested the hypothesis that the nature of these referents is increasingly inanimate. Since animacy is not available as a feature in the syntactically parsed or enriched texts, we semi-automatically added it to two of the enriched narrative texts: an OE one and a lModE one. This pilot experiment confirms the hypothesis, showing an increase in inanimate subjects from 37% in the OE text to 54% in the lModE one.

Since we have been arguing that one of the pressures for the change in subject functionality is the loss of clause-initial discourse-linking, we did a fourth experiment seeking to quantify the discourse-linking changes in the XP-S-V_{fin} environment (where XP can be a PP or an object NP). The results show a clear and steady decline of the proportion of clause-initial XPs being used for discourse-linking from approximately 60% in ME to 20% in lModE. The OE proportion (about 35%) deviates from this trend, but it is based on a small amount of data. Coding of further texts is needed to clarify whether this deviation is significant.

Our final point is that the four phenomena we studied in this article do not display identical patterns. This is not in accordance with the Constant Rate Effect, as proposed by e.g. Kroch (1989), but it should be remembered that this effect cannot be assumed to hold true for discourse phenomena or macro-structural planning, where the speaker is selecting one syntactic option from a range of many. This is why identifying diachronic trends in discourse requires even more data than identifying trends in diachronic syntax. Some constructions that can be argued to be primarily motivated by discourse or information-structural concerns, like passive infinitival clauses after verbs of thinking and declaring or locative inversion (e.g. Ward et al. 2002: 1365ff), also have a *metalinguistic* function in that they signal a particular text type and situate a text within a typology of discourse forms (see e.g. Fleischman 1990), which adds register and genre as complicating factors. This is a

domain where we only rarely find the patterns of straightforward competition that are the staple of diachronic syntax.[25] Although this means that our research cannot expect to uncover a direct statistical link between the decline of V2 and the rise of new discourse patterns, the patterns we found in this paper are nevertheless surprisingly consistent, particularly in view of the relatively small number of texts we have been able to enrich with referential information so far. They show that the research line we are taking, which involves combining syntactic information with referential states, is a promising one. We look forward to extending our experiments, in particular those that involve referential chains in the enriched texts that allow us to see how writers use the syntactic options at their disposal to help their readers keep track of referents.

5. Sources

The syntactically parsed English corpora that are currently being enriched are listed below, where the name of the corpus provides a link to the CoRD database:

[25] The position of the stressed-focus *it*-cleft in PDE is a case in point: as a new construction, it only shows partial overlap with the older inversion-structure whose decline may be argued to be responsible for its rise (cf (i) and (ii), from Los & Komen 2012):

(i) It was only after I had been in the room for a few minutes that I realized that everyone was staring at me
(ii) Only after I had been in the room for a few minutes did I realize that everyone was staring at me

Stressed-focus it-clefts have a range of other uses that do not show this overlap. Discourse and information structural functions are more reminiscent of the layering we see in grammaticalization than the competition we see in morphosyntactic change.

- YCOE: the York-Toronto-Helsinki Parsed Corpus of Old English Prose, which contains approximately 1.5 million words, divided over 100 texts (Taylor et al. 2003). Old English was around from 450 until 1150 A.D, but the earliest manuscripts are from the 9th century.
- PPCME2: the Penn-Helsinki Parsed Corpus of Middle English, second edition (Kroch & Taylor 2000). This corpus contains about 1.2 million words, which are divided over 55 text samples, and it covers a period from 1150 to 1500.
- PPCEME: the Penn-Helsinki Parsed Corpus of Early Modern English (Kroch et al. 2004). It contains about 1.7 million words, which are divided over 448 text samples. The period it covers runs from 1500 to 1710.
- PPCMBE: the Penn Parsed Corpus of Modern British English (Kroch et al. 2010). This corpus contains about 950.000 words, which are divided over 101 text samples, covering the period from 1700 until 1914.

6. References

Bech, Kristin. 2001. Word order patterns in Old and Middle English: a syntactic and pragmatic study. PhD dissertation, University of Bergen

Biber, Doublas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan & Finegan, Edward. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Carroll, Mary & Lambert, Monique. 2005. Reorganizing principles of information structure in advanced L2s. In *Educating for Advanced Foreign Language Capacities: Constructs, Curriculum, Instruction, Assessment*, H. Byrnes, H. Weger-Guntharp & K. Sprang (eds), 54-73. Washington: Georgetown: University Press.

Carroll, Mary, Rossdeutscher, Antje, Lambert, Monique & von Stutterheim, Christiane. 2008. Subordination in narratives and macrostructural planning: a comparative point of view. In *'Subordination' versus 'Coordination' in Sentence and Text: A Cross-linguistic Perspective*, Cathrine Fabricius-Hansen & Wiebke Ramm (eds), 161-184. Amsterdam; Philadelphia: John Benjamins.

Carroll, Mary, Stutterheim, Christiane von & Nuese, Ralph. 2004. The language and thought debate: A psycholinguistic approach. In *Multidisciplinary Approaches to Language Production*, Thomas Pechmann & Christopher Habel (eds), 183-218. Berlin: Mouton de Gruyter.

Downing, Angela & Locke, Philip. 2002. *A university Course in English Grammar*. London: Routledge.

Fischer, Olga, van Kemenade, Ans, Koopman, Willem & van der Wurff, Wim. 2000. *The Syntax of Early English*. Cambridge: Cambridge University Press.

Fleischman, Susan. 1990. *Tense and Narrativity: From Medieval Performance to Modern Fiction*. London: Routledge.

Garnham, Alan. 2001. *Mental Models and the Interpretation of Anaphora*. Hove, East Sussex: Psychology Press Ltd.

Halliday, Michael Alexander Kirkwood. 1994. *An Introduction to Functional Grammar*. London [etc.]: Edward Arnold.

Haug, Dag T. T., Jøhndal, Marius L., Eckhoff, Hanne M., Welo, Eirik, Hertzenberg, Mari J. B. & Müth, Angelika. 2009. Computational and linguistic issues in designing a syntactically annotated parallel corpus of Indo-European languages. *TAL* 50(2): 17-45. <http://www.atala.org/IMG/pdf/TAL-2009-50-2-01-Haug.pdf >.

Hinterhölzl, Ronald & van Kemenade, Ans. 2012. The interaction between syntax, information structure and prosody in word order change. In *The Oxford Handbook of the History of English*, Elizabeth Closs Traugott & Terttu Nevalainen (eds), 803-821. New York: Oxford University Press.

Hinterhölzl, Ronald & Petrova, Svetlana. 2010. From V1 to V2 in West Germanic. *Lingua* 120(2): 315-328.

Irmer, Matthias. 2011. *Bridging Inferences Constraining and Resolving Underspecification in Discourse Interpretation*. Berlin: Walter de Gruyter.

Johnson-Laird, P. N. 1983. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge, MA: Harvard University Press.

van Kemenade, Ans. 1987. *Syntactic Case and Morphological Case in the History of English*. Dordrecht: Foris Publications.

van Kemenade, Ans. 2012. Rethinking the loss of V2. In *The Oxford Handbook of the History of English*, Elizabeth Closs Traugott & Terttu Nevalainen (eds), 822-834. New York: Oxford University Press.

van Kemenade, Ans & Milicev, Tanja. 2012. Syntax and discourse in Old English and Middle English word order. In *Grammatical Change: Origins, Nature, Outcomes*. Dianne Jonas, Andrew Garrett & John Whitman (eds), 239-254. Oxford: Oxford University Press.

van Kemenade, Ans, Milicev, Tanja & Baayen, R. Harald. 2008. The balance between syntax and discourse in Old English. In *English Historical Linguistics 2006. Volume I: Syntax and Morphology*, Maurizio Gotti, Martina Dossena & Richard Dury (eds), 3-22. Amsterdam,Philadelphia: John Benjamins.

van Kemenade, Ans & Westergaard, Marit. 2012. Syntax and information structure: verb second variation in Middle English. In *Information Structure and Syntactic Change in the History of English*, Bettelou Los, María José López-Couso & Anneli Meurman-Solin (eds), 87-118. New York: Oxford University Press.

Komen, Erwin R. 2011. *Cesax: Coreference Editor for Syntactically Annotated XML Corpora.* Nijmegen, Netherlands: Radboud University Nijmegen, <http://erwinkomen.ruhosting.nl/software/Cesax > (7 November 2011).

Komen, Erwin R. 2012. Coreferenced corpora for information structure research. In *Outposts of Historical Corpus Linguistics: From the Helsinki Corpus to a Proliferation of Resources. [Studies in Variation, Contacts and Change in English 10]*. Jukka Tyrkkö, Matti Kilpiö, Terttu Nevalainen & Matti Rissanen (eds). Helsinki, Finland: Research Unit for Variation, Contacts, and Change in English, <http://www.helsinki.fi/varieng/journal/volumes/10/index.html > (24 November 2012).

Komen, Erwin R. 2013. Finding Focus: A Study of the Historical Development of Focus in English. PhD dissertation, Radboud University Nijmegen

Krifka, Manfred. 2007. Basic notions of information structure. In *Interdisciplinary Studies on Information Structure* 06, Caroline Féry, Gisbert Fanselow & Manfred Krifka (eds), 1-50.

Kroch, Anthony. 1989. Reflexes of grammar in patterns of language change. *Language Variation and Change* 1: 199-244.

Kroch, Anthony, Santorini, Beatrice & Diertani, Ariel. 2004. *Penn-Helsinki Parsed Corpus of Early Modern English*,

<http://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-2/index.html >.

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

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

Los, Bettelou. 2009. The consequences of the loss of verb-second in English: information structure and syntax in interaction. *English Language and Linguistics* 13(1): 97-125.

Los, Bettelou. 2012. The loss of verb-second and the switch from bounded to unbounded systems. In *Information Structure and Syntactic Change in the History of English*, Anneli Meurman-Solin, María José López-Couso & Bettelou Los (eds), 21-46. Oxford: Oxford University Press.

Los, Bettelou & Dreschler, Gea. 2012. The loss of local anchoring: From adverbial local anchors to permissive subjects. In *Rethinking Approaches to the History of English*, Terttu Nevalainen & Elizabeth Closs Traugott (eds), 859-872. New York: Oxford University Press.

Prince, Ellen. 1981. Toward a taxonomy of given-new information. In *Radical Pragmatics*, Peter Cole (ed), 223-255. New York: Academic Press.

Taylor, Ann, Warner, Athony, Pintzuk, Susan & Beths, Frank. 2003. *The York-Toronto-Helsinki Parsed Corpus of Old English Prose*, <http://www-users.york.ac.uk/~lang22/YCOE/YcoeHome.htm >.

Ward, Gregory, Birner, Betty & Huddleston, Rodney. 2002. Information packaging. In *The Cambridge Grammar of the English Language*, Rodney Huddleston & Geoffrey K. Pullum (eds), 1363-1448. Cambridge: Cambridge University Press.

Zwaan, Rolf A. & Radvansky, Gabriel A. 1998. Situation models in language comprehension and memory. *Psychological Bulletin* 123(2): 162-185.