Split coordination in English: Why we need parsed corpora

Ann Taylor & Susan Pintzuk, University of York

In this article we provide a practical demonstration of how syntactically annotated corpora can be used to investigate research questions with a diachronic depth and synchronic breadth that would not otherwise be possible. The phenomenon under investigation, split coordination, affects every type of coordinated constituent (subject/object DPs, predicate and attributive ADJPs, ADVPs, PPs, and DP objects of P) in Old English; and it, or a superficially similar construction, occurs continuously throughout the attested period from approx. 800 to the present day. We bring to bear different types of evidence to argue that split coordination in fact represents two different constructions, one of which remains stable over time while the other is lost in the post-Middle English period.

Dans cet article, nous fournissons une démonstration pratique de la façon dont les corpus syntaxiquement annotés peuvent être utilisés pour étudier des questions de recherche avec une profondeur diachronique et une largeur synchronique qui ne seraient pas possibles autrement. Le phénomène étudié, la coordination scindée, affecte tous les types de constituants coordonnés (DP sujet / objet, ADJP prédictif et attributif, ADVP, PP et objets DP de P) en vieil anglais; et elle, ou une construction superficiellement similaire, se produit continuellement tout au long de la période attestée d'environ. 800 à nos jours. Nous apportons différents types de preuves pour soutenir que la coordination scindée représente en fait deux constructions différentes, dont l'une reste stable dans le temps tandis que l'autre est perdue dans la période post-moyen-anglaise.

In diesem Artikel zeigen wir in einer praktischen Demonstration, wie syntaktisch annotierte Korpora verwendet werden können, um Forschungsfragen mit einer diachronen Tiefe und synchronen Breite zu untersuchen, die sonst nicht möglich wäre. Das untersuchte Phänomen, die Split-Koordination, betrifft jede Art von koordinierten Bestandteilen (Subjekt / Objekt-DPs, Prädikat- und attributive ADJPs, ADVPs, PPs und DP-Objekte von P) in Old English; und es, oder eine oberflächlich ähnliche Konstruktion, tritt ununterbrochen während der bezeugten Periode von ungefähr auf. 800 bis heute. Wir führen verschiedene Arten von Beweisen an, um zu argumentieren, dass die geteilte Koordination tatsächlich zwei verschiedene Konstruktionen darstellt, von denen eine im Laufe der Zeit stabil bleibt, während die andere in der Zeit nach dem mittleren Englisch verloren ist.

Key words: coordination, ellipsis, annotated corpora, syntactic change.

1. Introduction

In this article we provide a practical demonstration of how morpho-syntactically annotated (parsed) corpora, in particular the English Historical Parsed Corpora Series, can be used to investigate research questions with a diachronic depth and synchronic breadth that would not otherwise be possible. The English Historical Parsed Corpora Series is a collection of historical treebanks created at the University of Pennsylvania and the University of York, which provides continuous coverage of the English language from the earliest attested Old English (OE) texts through to Present-Day English (PDE). The corpora are all annotated using the same guidelines, so that syntactic variation and change can be tracked through the entire history of English.

The phenomenon under investigation, which we refer to descriptively as ‘split coordination’, is illustrated in 0, with examples from OE taken from the York-Toronto-Helsinki Corpus of Old English prose (YCOE). Every type of coordinated constituent (subject and object DPs, predicate and attributive ADJPs, ADVPs, PPs, and DP objects of P) can be split. Furthermore, this is not just an OE phenomenon; split coordination, or a construction that is

superficially similar, occurs continuously throughout the attested period from approximately 800 to the present day, as the parallel examples in 0 from PDE, taken from the Switchboard Corpus, demonstrate.

(1) a. DP subject
oðþæt ***þæt ad*** *wæs forburnen,* ***and ealle þa tunnan***
until the pile was burned and all the casks
"until the pile and all the casks were burned up"
(coaelive,+ALS_[Julian_and_Basilissa]:332.1143)1

b. DP object
God sende ða ***fyr*** *on merigen* ***and fulne swefel*** *him to*
God sent then fire in morning and foul brimstone him to
"God then sent fire and foul brimstone to him in the morning"
(coaelive,+ALS[Pr_Moses]:211.2976)

c. PP
& ***on sorhge*** *leofodon* ***&*** ***on geswincum*** *siþþan*
and in grief lived and in torment afterwards
"and [they] lived afterwards in grief and torment"
(colsigewZ,+ALet_4_[SigeweardZ]:117.49)

(2) a. *and this is where* ***my aunt*** *lives* ***and my uncle****,*
b. *you put, um,* ***really good vanilla flavoring*** *in it* ***and some butter***
c. *my only experience with it, I was* ***in Central America*** *for a while,* ***and, uh, in San Salvador, in El Salvador,***

Despite its synchronic range and diachronic persistence, split coordination has received surprisingly little attention in the diachronic literature. Its occurrence in OE is often mentioned (Kohonen 1978; Mitchell 1985: §§1464-72; Reszkiewicz 1966; Sielanko 1994; Traugott 1972); but beyond Perez Lorido's (2009) suggestive, but rather limited, study of split subjects in eight OE texts, it hasn't been seriously investigated. Its modern counterpart, which is most frequently analysed as a type of Gapping known as Stripping or Bare Argument Ellipsis (BAE), has been discussed in the literature since at least the sixties (Hankamer & Sag 1976; Johnson 2006; Reinhart 1991; Ross 1967), but no empirical corpus-based studies of its use exist.

One reason for the lack of quantitative, empirical investigations of this construction is, perhaps, that while the number of split coordinations is by no means negligible, neither is it high enough that sufficient numbers of examples can easily be collected without computational aids. Perez Lorido's study, based on manually collected data, is a case in point. He limits his study to coordinated subjects only, and the total number he collects from his eight texts is 731, of which 142 (19.4%) are split. By contrast, a search of the YCOE uncovers 3,391 coordinated subjects -- more than four times as many as Perez Lorido -- out of a total of 139,775 nominative subjects (2.4%), with 629 of these split (18.5%), not to mention over 2,000 cases of conjoined objects as well as smaller but still healthy numbers of the other categories. The situation in PDE is even more difficult, as here the construction occurs at frequencies well below 10%.

[1] Example references are to the corpus.

In addition to the problem of low frequency, this construction, whether split or not, is not uniquely marked by any lexical item; the only item common to these constructions is the conjunction itself. Therefore, without a parsed corpus, we could search only for *and* and other conjunctions and their variant spellings, or, in a part-of-speech (POS) tagged corpus, for the POS 'conjunction'. These retrievals will suffer badly from low precision (too much unwanted data), however, since the search will retrieve every token containing a conjunction, and there is no way to disregard conjunctions that are irrelevant, e.g. those conjoining clauses rather than smaller constituents, and no automatic way to separate split coordination from non-split coordination.

Given the low frequency and non-uniqueness of this construction, retrieving it manually from printed texts or even from a text/POS-tagged corpus will be at best limited in scope and inefficient, and at worst error-prone and unrepresentative.[3] In the English parsed corpora we use in this investigation, by contrast, (split) coordinations are explicitly marked, making it possible to quickly and accurately retrieve the relevant data.

2. The case study

As noted above, all coordinated categories can and do split, but to keep things manageable, we focus here only on subject coordination, the most common type. Further we limit the study to the following three research questions, out of the many we could pursue:

1. What is the frequency of split subject coordination over time? Is it a stable construction? Is it changing? In which direction?
2. Is the construction in the historical corpora the same in all respects to that found in PDE?
3. What factors (weight, information structure, etc.) affect splitting/non-splitting?

2.1 Extracting the data

The data for this study are taken from the following corpora:

Corpus	Size in words	Date range
The York-Toronto-Helsinki Corpus of Old English Prose (YCOE)	1,450,376	800-1150
Penn-Helsinki Parsed Corpus of Middle English 2 (PPCME2)	1,155,965	1150-1500
Parsed Corpus of Early English Correspondence (PCEEC)	2,159,132	1400-1710
Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME)	1,794,010	1500-1700
Penn-Helsinki Parsed Corpus of Modern British English (PPCMBE)	948,895	1700-1915
The Brown Corpus (Fiction and Imaginative Prose) (BROWN)	432,879	1960s
Wall Street Journal (WSJ)	851,496	1980s
CALLHOME (CH)	166,619	1990s
Switchboard (SWBD)	1,253,960	1990s

[3] This is not to say smaller manual studies are impossible, as Perez Lorido (2009) shows.

Although all the corpora are in ascii format and thus can be used on any platform and viewed and searched with any word processing program, in order to fully utilize the annotations, a search program sensitive to structure is required. We use CorpusSearch, a program conceived and designed by Ann Taylor & Anthony Kroch, and implemented by Beth Randall in Java. CorpusSearch is not corpus specific but will search any corpus in the correct format, including all the corpora in the English Parsed Corpora Series, and related corpora in other languages. Queries that can be used to extract the data for this study are included in Appendix II.

While all the corpora in our set are parsed in the Penn Treebank style, there are some differences between the historical and the present-day corpora with regard to how particular constructions, including coordination, are handled. For this reason, some of the searches differ in detail although the material retrieved is the same.

(3)a-b illustrates the structure of split coordination in the historical corpora. The 2nd conjunct is linked to the rest of the subject by means of a co-indexed trace (*ICH*).[4] Example (3a) in tree form is given in (3c).

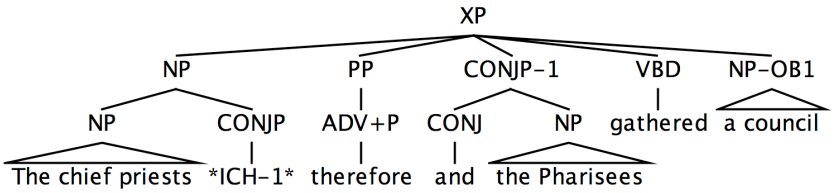
(3) The structure of split coordination in the historical corpora.

a.
```
/~*
The chief priests therefore and the Pharisees gathered a council,
(ERV-NEW-1881,11,40J.1030)
*~/
( (IP-MAT (NP-SBJ (NP (D The) (ADJ chief) (NS priests)) <-- 1st conjunct
                      (CONJP *ICH*-1))    <-- trace of 2nd conjunct
          (PP (ADV+P therefore))
          (CONJP-1 (CONJ and)             <-- 2nd conjunct
                   (NP (D the) (NPRS Pharisees)))
          (VBD gathered)
          (NP-OB1 (D a) (N council))
          (. ,))
  (ID ERV-NEW-1881,11,40J.1030))
```

b.
```
/~*
Besides both Jesus was invited, and his Disciples to the Marriage.
(PURVER-NEW-1764,2,1J.98)
*~/
( (IP-MAT (ADVP (ADV Besides))
          (NP-SBJ (CONJ both)       <-- 1st conjunct
                  (NP (NPR Jesus))
                  (CONJP *ICH*-1))  <-- trace of 2nd conjunct
          (BED was)
          (VAN invited)
          (, ,)
          (CONJP-1 (CONJ and)       <-- 2nd conjunct
                   (NP (PRO$ his) (NS Disciples)))
          (PP (P to)
              (NP (D the) (N Marriage)))
          (. .))
  (ID PURVER-NEW-1764,2,1J.98))
```

[4] ICH is an inherited label from the Penn Treebank. It is not a theoretical construct, but just stands for 'Interpret Constituent Here'.

c. Example (a) in tree form



In the PDE corpora the label dominating the 2nd conjunct is different (NAC rather than CONJP)[5], but the structure is essentially the same, as can be seen in 0, and thus retrieval is equally easy and accurate.

(4) The structure of split coordination in the PDE corpora (SWBD)

```
/~*
but then today the wind has dropped off, and also, the temperature,
*~/
( (S (CONJP (CC but) (RB then))
     (NP-TMP (NN today))
     (NP-SBJ (NP (DT the) (NN wind))      <-- 1st conjunct
             (NAC (-NONE- *ICH*-1)))      <-- trace of 2nd conjunct
     (VP (VBZ has)
         (VP (VBN dropped)
             (PRT (RP off))
             (, ,)
             (NAC-1 (CC and)              <-- 2nd conjunct
                    (ADVP (RB also))
                    (, ,)
                    (NP (DT the) (NN temperature)))))
```

In addition to the split coordinations, we need to collect the non-split cases, not only in order to generate frequencies, but also to act as a control on any potential explanation for why splitting occurs. Here the difference between the parsing of the historical and PDE corpora is a bit larger, but the relevant examples can, nevertheless, be retrieved with great accuracy in both cases. Examples of non-split coordinations in the historical corpora and the PDE corpora are given in 0 and 0.

[5] NAC stands for 'Not A Constituent' and results from the lack of a Conjunction Phrase encompassing the conjunction and 2nd conjunct in the Penn Treebank (PTB) parsing scheme; see example 0. Note that no label (-NONE-) is given to the trace in 0; in (3)a-b, the trace is a CONJP. Split coordination does not occur in the Wall Street Journal, the first Penn Treebank corpus that was parsed and the one that much of the parsing scheme was developed to handle. The NAC label, originally used for other constructions, was co-opted to handle this construction during the parsing of the spoken Switchboard corpus.

(5) The structure of non-split coordinations in the historical corpora

```
/~*
for the bagpipes and the musicke went to wracke - (ARMIN,-E2-H:11.98)
*~/
( (IP-MAT (CONJ for)
          (NP-SBJ (NP (D the) (N+NS bagpipes))        <-- 1st conjunct
                  (CONJP (CONJ and)
                         (NP (D the) (N musicke))))   <-- 2nd conjunct
          (VBD went)
          (PP (P to)
              (NP (N wracke)))
          (. -))
  (ID ARMIN,-E2-H:11.98))
```

(6) The structure of non-split coordination in PDE corpora (SWBD)

```
/~*
Both my mother 's parents and my father 's parents were immigrants E_S
*~/
( (S (NP-SBJ (DT Both)
             (NP (NP (PRP$ my) (NN mother) (POS 's))
                 (NNS parents))                       <-- 1st conjunct
             (CC and)
             (NP (NP (PRP$ my) (NN father) (POS 's))
                 (NNS parents)))                      <-- 2nd conjunct
     (VP (VBD were)
         (NP-PRD (NNS immigrants)))
     (-DFL- E_S))
```

2.2 The distribution of split subject coordination over time

The first step is to retrieve all the relevant tokens from each corpus, i.e. all coordinated subjects, whether they are split or not (see Appendix II, A.1). We can then separate the split from the non-split tokens (Appendix II, A.2). From these data, we can get an overall picture of the distribution of split subject coordination from the OE to PDE periods in all the corpora we have available, ordered approximately by date, as shown in Table 1.

Table 1. The frequency of split coordination in English historical and present-day corpora.

Corpus	Non-split	Split	Total	%split
YCOE (800-1150)	2762	629	3391	18.55%
PPCME (1150-1500)	2212	277	2489	11.13%
PCEEC (1400-1710)	4049	292	4341	6.73%
PPCEME (1500-1700)	3948	223	4171	5.35%
PPCMBE (1700-1915)	1922	14	1936	0.72%
BROWN (1960s)	696	7	703	1.00%
WSJ (1980s)	1633	0	1633	0.00%
CALLHOME (1990s)	60	5	65	7.69%
SWBD (1990s)	398	38	436	8.72%

Table 1 reveals a basic downward trend in the frequency of split subject coordination over time, with a strong rise at the end of the 1990s. This rise, however, is somewhat deceiving, as it comes from two speech corpora. Clearly this construction, at least in PDE, is restricted to more oral registers. The best modern comparators to our earlier corpora, all of which are necessarily written, are thus the written BROWN/WSJ corpora.[6] Given the oral aspect of this construction, we also need to be careful with the PCEEC corpus, which, while obviously not representing speech, is made up solely of personal letters and thus is designed to be as vernacular as possible. The PCEEC also overlaps partially in time with the PPCME and PPCEME, further complicating matters. We will thus exclude the PCEEC as well as the speech corpora from the main investigation. Removing the PCEEC, CALLHOME and SWBD from Table 1, and collapsing BROWN and WSJ, gives Table 2.

Table 2. The frequency of split coordination in English historical and present-day corpora, excluding and collapsing corpora.

Corpus	Non-split	Split	Total	%split
YCOE (800-1150)	2762	629	3391	18.55%
PPCME (1150-1500)	2212	277	2489	11.13%
PPCEME (1500-1700)	3948	223	4171	5.35%
PPCMBE (1700-1915)	1922	14	1936	0.72%
WSJ/BROWN (late 20th c.)	2329	7	2336	0.30%

We have now answered our first question. Split subject coordination has always been a low frequency construction (in written texts), but shows a clear and fairly steady decrease over time. In addition, we have also identified one issue relevant to question 3, that split coordination is apparently sensitive to register.

2.3 A comparison of PDE with earlier stages of the language

We focus in this section on the cause of the decline in frequency of the split coordination construction. The sensitivity to register raises the possibility that the decline evident in Table 2 is simply an external effect, perhaps the result of standardization and/or a prescription against this construction in writing, and doesn't represent a change in the syntax of the language, but only a change in register norms, or something similar. This might explain the fact that the major decline in split coordination post-dates the Middle English period, and that in the modern language it can be found in speech and fiction; in contrast, the Wall Street Journal corpus furnishes no examples despite its large size. A closer look at the data, however, raises the possibility that, although the OE and PDE constructions appear superficially similar, we are actually looking at two different constructions. The relevant difference can be seen by comparing the examples in 0 and 0 and in 0 and 0; while the 2nd conjunct in PDE is overwhelmingly found in clause-final position,[7] the same is not true in the earlier stages of the language. In fact, final position for the 2nd conjunct in Old and Middle

[6] This is the best available, but clearly not perfect, as the range of registers represented is much more limited than in the earlier corpora. A much better comparator would be the written part of the BNC, but as it is not parsed, it is impossible to extract the relevant data, as discussed above.

[7] Two non-final cases occur in SWBD, given in (i) and (ii). The first case has an indirect question following the 2nd conjunct; and the second case, a relative clause modifying both conjuncts.

(i) on publicity and letting **realtors** know **and key people** how wonderful the schools are

(ii) I hate to see **a car** going down the street, **or even a truck or bus for that matter**, that 's putting out a lot of dark smoke,

English is actually less common than non-final position: only about 30% of the subject 2nd conjuncts occur in final position in these early stages.

(7) a. *but then today **the wind** has dropped off, and also, the **temperature**,* (SWBD)
b. ***a new carrier** was coming in **and, uh, the, uh, attendant, uh, support vessels.*** (SWBD)
c. ***A cold supper** was ordered **and a bottle of port.*** (BROWN)
d. *"**Fear** possessed me, **and the certainty of war**", he has related.* (BROWN)

(8) a. ***Hys apostoli** arærdon **and heora æftergengan** manega menn of deaðe*
his apostles raised and their followers many men from death
"His apostles and their followers raised many men from death"
(YCOE: coaelhom,+AHom_6:324.1027)

b. *But so it bifel þat **Rudak** was slayn, **and Skater also**, in pleyn bataile*
but so it befell that Rudak was slain and Skater also in open combat
"so it befell that Rudak and Skater also were slain in open combat"
(PPCME: CMBRUT3,23.691)

c. *And **both Iesus** was called, **and his disciples**, to the mariage*
(PPCEME: AUTHNEW,-E2-H:II,1J.166)

Given the overwhelming final position of 2nd conjuncts in PDE, modern syntactic accounts of this phenomenon have analysed it as a kind of ellipsis, referred to as Bare Argument Ellipsis (BAE) or Stripping in the literature (Hankamer & Sag 1976; Johnson 2006; Reinhart 1991; Ross 1967). Thus, example 0d is derived as in 0 from two full conjoined clauses with deletion under identity of everything in the second clause except the subject.[8]

(9) **Fear** possessed me, **and the certainty of war** ~~possessed me~~ (BROWN)

If the derivation involves movement of the remnant (i.e., the XP of the 2nd conjunct) to a left peripheral clause position prior to deletion under identity of the remainder of the second clause (Johnson 2006: 425; cf. also Busquets 2006; Konietzko & Winkler 2010), then it accounts as well for cases where objects and other coordinated constituents are split, as shown in 0:

(10) you put, um, **really good vanilla flavoring** in it and you put **some butter** in it
you put, um, **really good vanilla flavoring** in it **and some butter-i** you put t-i in it
you put, um, **really good vanilla flavoring** in it **and some butter-i** ~~you put t-i in it~~

This analysis necessarily produces a clause-final 2nd conjunct, and thus works well for PDE, but it is not so clear how the examples in 0 could be derived by the same mechanism, since the 2nd conjunct is clause-internal, followed by a direct object in 0a, a locative PP in 0b and a PP complement of the verb in 0c. Clearly, earlier stages of English have another way of deriving split coordinations that can be used instead of, or in addition to, BAE. Given this, the proportion of final to non-final 2nd conjuncts over time is clearly of interest. Our next set of searches, therefore, takes all the cases of split subject coordinations and divides them into

[8] Another possible way to derive these examples is by extraposition of the 2nd conjunct (Munn 1993), which may be more or less attractive depending on your theory and the structure it assigns to coordinated phrases. We will not pursue this alternative here.

cases with a final or non-final 2nd conjunct. Of course, we could simply write a query to extract each type from each corpus individually; there is, however, a more efficient way, one that in addition prepares for subsequent searches of the data. CorpusSearch includes a facility to code tokens for any feature for which it is possible to search. We can therefore take our set of split coordinated subjects and code them for the position of the 2nd conjunct. (Appendix II, A.3). The data can then be exported and analysed in a spreadsheet or statistical analysis program, such as R. Examples are given in 0 for two tokens, the first with a final 2nd conjunct, indicated by the CODING node, the second with a non-final 2nd conjunct.

(11) Using coding strings for easier calculation of statistics

a.
```
( (IP-SUB (CODING final)
          (CS1-NP-NOM^1 (NP-NOM (D^N +t+at) (N^N ad))
                        (CONJP *ICH*-1))
          (BEDI w+as)
          (VBN forburnen)
          (, ,)
          (CONJP-1 (CONJ and)
                   (NP-NOM (Q^N ealle) (D^N +ta) (N^N tunnan))))
  (ID coaelive,+ALS_[Julian_and_Basilissa]:332.1143))
```

b.
```
( (IP-MAT-SPE (CODING non.final)
              (NEG+CONJ ne)
              (ADVP-LOC (ADV^L +t+ar))
              (CS1-NP-NOM^1 (NP-NOM (N^N w+adla))
                            (CONJP *ICH*-1))
              (NEG ne)
              (BEPI bi+d)
              (, ,)
              (CONJP-1 (NEG+CONJ ne)
                       (NP-NOM (ADJ^N wanhal)))
              (VBN gemet)
              (. .))
  (ID coaelive,+ALS_[Thomas]:80.7594))
```

Table 3 shows the results of separating 2nd conjuncts by position: there is a steady decline in non-final 2nd conjuncts, with an unexpected peak in the PPCMBE.

Table 3. Split coordinated subjects: position of 2nd conjunct.

Corpus	Final	Non-final	Total	%non-final
YCOE (800-1150)	419	210	629	33.4%
PPCME (1150-1500)	202	75	277	27.1%
PPCEME (1500-1700)	179	44	223	19.7%
PPCMBE (1700-1915)	6	8	14	57.1%
WSJ/BROWN (late 20th c.)	6	0	6	0.0%

While it is not possible to investigate this spike in detail, due to space restrictions, a quick look at the non-final examples shows that four out of eight are from the Bible and repeat the word order of an earlier version, as illustrated in 0. Five out of the eight (three from the Bible) have a subject split only by a discourse particle (*therefore, then, too*, etc.), which calls

into question their evidence for syntactically split constituents. The remaining two examples[9] of this type in the PPCMBE give scant evidence for the continuation of this construction post-1700, and we can thus safely date the loss of this type of splitting to the end of the EME period.

(12) a. LME (1764)
*Besides **both Jesus** was invited, **and his Disciples** to the Marriage.*
(PPCMBE: PURVER-NEW-1764,2,1J.98)

b. EME (1611)
*And **both Iesus** was called, **and his disciples**, to the mariage.*
(PPCEME: AUTHNEW,-E2-H:II,1J.166)

(13) a. LME (1764)
***The real benefits** then **which have been conferred on us by the Resurrection of our Lord, the substantial advantages which it has effected for us in our state of religious probation**, seem to be the two following.*
(PPCMBE: FROUDE-1830,2,50.347)

b. LME (1905)
***Small cutters**, too, **or centre-boards, handled by local amateurs**, will now and again come dashing out…*
(PPCMBE: BRADLEY-1905,201.46)

2.4 Factors favouring the splitting of conjuncts

Turning finally to our third research question, what factors trigger the splitting of coordinated subjects, we can use the corpora to investigate at least one possible factor: weight (or length or complexity). Mitchell (1985: §§1464-72) subsumes split coordinations under a process he calls 'splitting of heavy groups' and it has been shown that weight is a key factor in rightward movement processes in OE in general (Pintzuk & Taylor 2006; Taylor & Pintzuk 2011, 2012a, 2012b, 2014). Thus, despite Perez Lorido's claim that weight is not a factor in the case of split subject coordination,[10] it seems a good candidate. The automatic counting of words and/or nodes in parsed corpora is possible with CorpusSearch, and taking advantage of the coding feature discussed above, measuring weight in terms of number of words[11] is generally quite straightforward. Here we test two hypotheses: (1) longer coordinations are more likely to split than shorter ones (based on Mitchell's heavy groups claim); and (2) the weight of the 2nd conjunct (possibly in comparison with the weight of the 1st conjunct) is a factor in promoting splitting, i.e., heavier 2nd conjuncts are more likely to split, *pace* Perez Lorido.

---

[9] One example is ***Paleness** sits on every face; **confused tremor and fremescence**; waxing into thunder-peals, of Fury stirred on by Fear*. (PPCMBE: CARLYLE-1837,1,149.338), in which the non-finite clause *waxing…* could be taken as belonging to *confused tremor and fremescence*, in which case the 2nd conjunct is final. The second example, *Was not **both my Topsail Yards** wounded, **and Maintop-Mast**, when I then bore down to the Enemy?* (PPCMBE: HOLMES-TRIAL-1749,41.654) is taken from trial data and represents direct speech. It is similar to examples found in the spoken Switchboard Corpus, as noted in footnote 6.

[10] Perez Lorido's numbers show that on average, a split 2nd conjunct is one word longer than a non-split one, but he assumes without testing that this difference is insignificant.

[11] Weight/length/complexity can be measured in various ways (cf. Taylor & Pintzuk 2012b and references therein). In practice, weight is such a robust effect that it makes little or no difference what measure is used. Since word count is easy to automate, we use number of words here as the measure of weight.

CorpusSearch counts words within a given node. As non-split coordinations are dominated by a single node (cf. example 0), obtaining the number of words in a non-split coordination is completely straightforward. Split coordinations, on the other hand, are not dominated by a single node (cf. example (3)) with the result that each conjunct must be counted separately and the results summed. For technical reasons related to how CorpusSearch works, it is not easily possible to calculate the length in words of coordinations including embedded clauses (e.g. relative clauses) or for coordinations with shared constituents, as in *the husband and wife* or *the rude savage or uncultured boor*, where the determiner is shared in both cases; these two constructions are thus excluded from the statistics in Tables 4-5 (Appendix II, A.5).

Table 4 shows the average length of split and non-split coordinations across the three early corpora; the later corpora are omitted as the splitting of coordinations in written texts is essentially over by 1700 (Table 2). In each case the average length of split coordinations is longer than that of non-split coordinations. The extra length is small (about 0.5-1.0 word longer), but significant (at $p < 0.05$) for the first two corpora. For the PPCEME, the difference is not significant ($p < 0.1$). This is a potentially interesting difference, but a full exploration is beyond the scope of this paper.

Table 4. Average total length of coordination in words.

Corpus	Split	Non-split	Difference
YCOE (800-1150)	6.88	5.99	0.89
PPCME (1150-1500)	7.25	6.46	0.79
PPCEME (1500-1700)	7.96	7.29	0.67
Average[12]	7.37	6.58	0.78

It is certainly not the case that only heavy groups split, as the split three-word examples in 0 show. In our data, three-word coordination split about 5% of the time, while coordinations of length four words and above split on average about 15% of the time, showing no particular trend as length increases, as shown in Table 5. These data appear to call into question the traditional labelling of this phenomena as 'splitting of heavy groups' (as also noted by Perez Lorido 2009: 35).

(14) a. ***Adam*** *þagyt* ***&*** ***Eua*** *næron onlysde,*
Adam yet and Eve not-were liberated
"Adam and Eve were not yet liberated"
(YCOE: coblick,HomS_26_[BlHom_7]:87.88.1110)

b. *þan shulde* ***pees*** *haue bene,* ***and reste*** *amongus ham, wiþouten eny envy.*
then should peace have been and rest among us without any envy
"then there should have been peace and rest among us without any envy"
(PPCME: CMBRUT3,220.3966)

c. *Did* ***he*** *pull down the Hay* ***or you****?*
(PPCEME: LISLE,-E3-H:IV,114C2.104)

[12] The slight discrepancy here is due to rounding errors.

Table 5. Percentage of split coordinations by total number of words.

Total length in words	N	%split
3	1399	4.9%
4	753	15.3%
5	1571	13.6%
6	751	16.0%
7	620	13.4%
8	416	18.0%
9	341	18.8%
10	157	14.6%

The second question, regarding the length of the 2nd conjunct in particular, is slightly more difficult to test due to the way that coordinations are annotated in the corpora. Most coordinations have the structure illustrated in 0, repeated here as 0, and thus counting the length of the 2nd conjunct, which is entirely dominated by the CONJP node, is straightforward. This number can then be subtracted from the total giving the length of the 1st conjunct as well. Slightly problematic here are coordinations which consist of conjoined single words. These coordinations are annotated as flat structures, i.e., without a CONJP node, as illustrated in 0.[13] In these cases, there is no defined 2nd conjunct that can be counted, and thus the counting of this category has to be done by hand. As in these cases the 1st conjunct is necessarily one word long, however, this type can be counted in the same way as other unsplit coordinations.

(15)

```
/~*
for the bagpipes and the musicke went to wracke - (ARMIN,-E2-H:11.98)
*~/
( (IP-MAT (CONJ for)
          (NP-SBJ (NP (D the) (N+NS bagpipes))  <-- structured coordination
                  (CONJP (CONJ and)
                         (NP (D the) (N musicke))))
          (VBD went)
          (PP (P to)
              (NP (N wracke)))
          (. -))
  (ID ARMIN,-E2-H:11.98))
```

[13] This approach to annotating single word coordinations was adopted wholesale from the Penn Treebank in order to save time and effort in the annotation process. In retrospect, it was clearly a mistake not to annotate all coordinations in a consistent manner.

(16)

```
/~*
But error and phantasie, do commonlie occupie, the place of troth and
iudgement. (ASCH,-E1-P2:14V.94)
*~/

( (IP-MAT (CONJ But)
          (NP-SBJ (N error) (CONJ and) (N phantasie)) < "flat" coordination
          (, ,)
          (DOP do)
          (ADVP (ADV commonlie))
          (VB occupie)
          (, ,)
          (NP-OB1 (D the)
                  (N place)
                  (PP (P of)
                      (NP (N troth) (CONJ and) (N iudgement))))
          (. .))
  (ID ASCH,-E1-P2:14V.94))
```

A final difficulty is the structure of coordinations with more than two conjuncts. In the corpora, these have the structure given in 0. Two questions arise here, one conceptual and one practical. Conceptually, we need to decide what counts as the 2nd conjunct in non-split coordinations. If we look at the split cases of coordinations with multiple conjuncts, it is clear that the split overwhelmingly occurs after the 1st conjunct. Thus, we should count everything except the 1st conjunct together. This approach, however, leads to a practical problem, because these conjuncts are not dominated by a single node in the annotation and thus can't be counted automatically. In this case, therefore, we do the opposite of what we did with binary coordinations. We count the 1st conjunct (which is dominated by a node) and subtract it from the total, giving the length of the 2nd conjunct.

(17)

```
/~*
And the Lord sayd vnto Aaron, Thou and thy sonnes, and thy fathers house
with thee, shall beare the iniquitie of the Sanctuary: (AUTHOLD,-E2-
P1:XVIII,1N.1125)
*~/
( (IP-MAT-SPE (NP-SBJ (NP (PRO Thou))
                      (CONJP (CONJ and)
                             (NP (PRO$ thy) (NS sonnes)))
                      (, ,)
                      (CONJP (CONJ and)
                             (NP (NP-POS (PRO$ thy) (N$ fathers))
                                 (N house)
                                 (PP (P with)
                                     (NP (PRO thee))))))
              (, ,)
              (MD shall)
              (VB beare)
              (NP-OB1 (D the)
                      (N iniquitie)
                      (PP (P of)
                          (NP (D the) (N Sanctuary)))))
  (ID AUTHOLD,-E2-P1:XVIII,1N.1125))
```

With respect to the length of the 2nd conjunct as a factor in favouring splitting, our results confirm Perez Lorido's, as shown in Table 6:[14] on average the 2nd conjunct in a split coordination is one half to one word longer than in a non-split coordination. As with the overall length, this difference is significant ($p < 0.01$) for the two earlier corpora, but not for the PPCEME. By contrast the difference in the length of the 1st conjunct between split and non-split coordinations is much smaller, varies in direction and only the average difference over all three corpora is significant.[15]

Table 6. Average length of the 2nd and 1st conjunct and difference (split - non-split).

Corpus	2nd conjunct			1st conjunct		
	Split	Non-split	Difference	Split	Non-split	Difference
YCOE	5.06	4.08	0.98	1.82	1.92	-0.10
PPCME	5.22	4.35	0.87	2.03	2.12	-0.09
PPCEME	5.51	4.93	0.58	2.46	2.36	0.10
Average	5.18	4.46	0.72	2.00	2.13	-0.13

If we look at the percentage of split coordinations by the length of the 2nd conjunct (Table 7), we see that most of the effect is concentrated at the low end, with the percentage rising between 2 and 4 words,[16] but then more or less leveling off, in the same way as in Table 5.

Table 7. Percentage of split coordinations by length of 2nd conjunct.

Length of 2nd conjunct in words	N	%split
2	1620	5.1%
3	2047	12.5%
4	1062	17.1%
5	621	16.1%
6	451	19.7%
7	273	19.0%
8	191	15.2%
9	134	15.7%
10	81	11.1%

Thus, length (weight/complexity), particularly of the 2nd conjunct, is clearly a factor in the splitting of coordinations, but the effect appears to be fairly small and makes the most difference for the shortest items. As usual with this factor, it is not clear what it represents; for some discussion of this issue, see Arnold et al. (2000) ; Taylor & Pintzuk (2012b).

[14] Conjunctions occurring before the 1st conjunct are counted as part of that conjunct; conjunctions occurring before the 2nd conjunct are counted as part of that conjunct.

[15] Welch 2 sample t-test: 1st conjunct: YCOE: p = 0.09, PPCME: p = 0.40, PPCEME: p = 0.60, average p = 0.02; 2nd conjunct: YCOE: p = 3.252e-08, PPCME:0.002, PPCEME: p = 0.07, average p = 1.177e-07)

[16] The coordination is included as part of the count of the 2nd conjunct, thus a two-word 2nd conjunct is generally made up of a conjunction plus a single word conjunct.

Another factor that is likely to play a role in split coordinations is information structure (IS), which is well-known to influence rightward movement (e.g., Hinterhölzl 2009; Pintzuk & Taylor 2006). Perez Lorido claims that a split 2nd conjunct is defocused, while when non-split, it is focused or foregrounded and generally receives more "communicative attention" (2009: 42ff).[17] Kiss (1996) for PDE and Biberauer & Kemenade (2011) for OE/ME propose two subject positions, the higher of which is reserved for specific subjects and the lower for non-specific. Although we would not claim that the conjuncts in split subject coordinations always fill the two subject positions,[18] specificity as a factor in leftward movement of the 1st conjunct is a plausible hypothesis. Finally, it is well known that earlier positions in the clause favour given information and later ones new (see, for example, the 'Given Before New Principle' of Gundel (1988)), and thus IS may also be relevant to this construction. Unfortunately, the annotation of IS is much more difficult and time-consuming than the annotation of syntactic structure, as well as far less advanced; IS is not included in the annotation of the English Historical Parsed Corpora Series. As a result most studies which include an IS component up to now have been done manually (Bech 2001; Taylor & Pintzuk 2011, 2012a, 2012b, 2014). Other corpus projects (e.g., PROIEL, ISWOC) have started to explore methods to annotate information status along with syntax in their corpora, and in future the spread and ease of carrying out such investigations should increase.

3 Conclusion

In this paper we have demonstrated, via a case study of split coordination, how researchers can track a construction across a long time period and investigate possible hypotheses concerning the frequency of occurrence of the construction over time, its syntactic structure and the factors that influence its use in different contexts. Like many syntactic constructions, split coordinations are extremely difficult to extract without a parsed corpus, since the lexical items and parts of speech that it contains are not limited to phrasal coordinations. In addition, as a low frequency construction, particularly in the later periods, the effort and time needed to find and extract the relevant examples would be difficult to justify. Bringing to bear different types of evidence (structural, discourse/performance effects, rate of change, etc.) we test the hypothesis that split coordination in fact represents two different constructions, one of which, Bare Argument Ellipsis, remains stable over time, while the other – which involves the movement of the 2nd conjunct out of the conjoined phrase -- is lost in the post-Middle English period. As a discussion of the latter construction goes beyond the scope of this paper, the interested reader is referred to Taylor & Pintzuk (2017) for an analysis.

References

Arnold, Jennifer, Anthony Losongco, Thomas Wasow & Ryan Ginstrom. 2000. Heaviness vs. newness: The effects of structural complexity and discourse status on constituent ordering. *Language* 76(1). 28-55.

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

Biberauer, Theresa & Ans van Kemenade. 2011. Subject positions and information-structural diversification in the history of English. *Catalan Journal of Linguistics* 10. 17-69.

Busquets, Joan. 2006. Stripping vs. VP-ellipsis in Catalan: What is deleted and when? *Probus*

[17] We do not find Perez Lorido's analysis convincing due to the lack of any objective measure of the differences he claims in the IS of split and non-split 2nd conjuncts. However, this does not negate the possibility, indeed the strong probability, that IS is involved at some level.

[18] Both subject positions precede the position normally filled by the finite verb (T or equivalent). In some cases both conjuncts do precede T, as in (3)a and 0a, and thus likely occupy the two subject positions (cf. Kemenade & Milićev 2012: 249). However, frequently only the 1st conjunct precedes the finite verb.

18. 159-187.
Gundel, Jeanette. 1988. Universals of topic-comment structure. In Michael Hammond, Edith Moravcsik & Jessica Wirth (eds.), *Studies in syntactic typology,* 209-239. Amsterdam: John Benjamins.
Hankamer, Jorge & Ivan Sag. 1976. Deep and surface anaphora. *Linguistic Inquiry* 7. 391-426.
Hinterhölzl, Roland. 2009. Information structure and unmarked word order in (Older) Germanic. In Caroline Féry & Malte Zimmerman (eds.), *Information Structure from Different Perspectives,* 282-304. Oxford: Oxford University Press,.
Johnson, Kyle. 2006. Gapping. In Martin Everaert & Henk van Riemsdijk (eds.), *The Blackwell companion to syntax*, 407-435. Oxford: Blackwell.
Kiss, Katalin. 1996. Two subject positions in English. *Linguistic Review* 13. 119-142.
Kemenade Ans van, & Tanja Milićev. 2012. Syntax and discourse in Old English and Middle English word order. In Dianne Jonas & Stephen Anderson (eds.), *Grammatical change: Origins, nature, outcomes: Proceedings of DIGS VIII*, 237-255. Oxford: Oxford University Press.
Kohonen, Viljo. 1978. *On the development of English word order in religious prose around 1000 and 1200 AD*. Åbo: Åbo Akademi Foundation.
Konietzko, Andreas & Susanne Winkler. 2010. Contrastive ellipsis: Mapping between syntax and information structure. *Lingua* 120. 1436-1457.
Mitchell, Bruce. 1985. *Old English syntax*. Oxford: Oxford University Press.
Perez Lorido, Rodrigo. 2009. Reconsidering the role of syntactic "heaviness" in Old English split coordination. *Studia Anglica Posnaniensia* 45. 31-56.
Pintzuk, Susan & Ann Taylor. (2006) The loss of OV order in the history of English. In Ans van Kemenade & Bettelou Los (eds.), *The handbook of the history of English,* 249-278. Oxford: Blackwell.
Reinhart, Tanya. 1991. Elliptic conjunctions - Non-quantificational LF. In Kasher, Aka (ed.), *The Chomskyan turn*, 360-384. Oxford: Blackwell.
Reszkiewicz, Alfred. 1966. Split constructions in Old English. In Mieczyslaw Brahmen, Stanislaw Helsztyński & Julian Krzyzanowski (eds.), *Studies in language and literature in honour of Margaret Schlauch*, 313-326. Warsaw: Polish Scientific Publishers.
Ross, John R. 1967. *Constraints on variables in syntax*. PhD thesis, MIT.
Sielanko, Elzbieta. 1994. Split coordinated structures in late Old English. *Studia Anglica Posnaniensia* 24. 58-72.
Taylor, Ann & Susan Pintzuk. 2011. The interaction of syntactic change and information status effects in the change from OV to VO in English. *Catalan Journal of Linguistics* 10. 71-94.
Taylor, Ann & Susan Pintzuk. 2012a. The effect of information structure on object position in Old English: A pilot study. In Maria-Jose López-Couso, Bettelou Los & Anneli Meurman-Solin (eds.), *Information structure and syntactic change*, 47-65. Oxford: Oxford University Press.
Taylor, Ann & Susan Pintzuk. 2012b. Rethinking the OV/VO alternation in Old English: The effect of complexity, grammatical weight and information structure. In Terttu Nevalainen & Elizabeth Traugott (eds.), *The Oxford handbook of the history of English,* 835-845. Oxford: Oxford University Press.
Taylor, Ann & Susan Pintzuk. 2014. Testing the theory: Information structure in Old English. In Kristin Bech & Kristine G. Eide (eds.) *Information structure and syntactic change in Germanic,* 53-77. Amsterdam: John Benjamins.
Taylor, Ann & Susan Pintzuk. 2017. Split coordination in Early English. In Bettelou Los & Pieter de Haan (eds.) *Word order change in acquisition and language contact: Essays in*

honour of Ans van Kemenade, 155-183. Amsterdam: John Benjamins.
Traugott, Elizabeth. 1972. *A history of English syntax*. New York: Holt, Rinehart, and Winston.
Vicente, Luis. 2010. On the syntax of adversative coordination. *Natural Language and Linguistic Theory* 28. 381-415.
Visser, Fredericus. 1963-73. *An historical syntax of the English language*. Leiden: Brill.

Ann Taylor
Department of Language and Linguistic Science
University of York
Heslington, York YO10 5DD
United Kingdom
ann.taylor@york.ac.uk

Susan Pintzuk
Department of Language and Linguistic Science
University of York
Heslington, York YO10 5DD
United Kingdom
susan.pintzuk@york.ac.uk

Appendix I: Corpora

CallHome Weischedel, Ralph, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El- Bachouti, Robert Belvin & Ann Houston. OntoNotes Release 5.0 LDC2013T19. Web Download. Philadelphia: Linguistic Data Consortium, 2013.

Brown Corpus, Switchboard, Wall Street Journal, Marcus, Mitchell, Beatrice Santorini, Mary Ann Marcinkiewicz, & Ann Taylor. Treebank-3 LDC99T42. Web Download. Philadelphia: Linguistic Data Consortium, 1999.

PROIEL Pragmatic Resources in Old Indo-European Languages: http://www.hf.uio.no/ifikk/english/research/projects/proiel

ISWOC Information Structure and Word Order Change in Germanic and Romance Languages: http://www.hf.uio.no/ilos/english/research/projects/iswoc/

The English Historical Parsed Corpora Series:

YCOE: Taylor, Ann, Anthony Warner, Susan Pintzuk & Frank Beths. 2003. York-Toronto-Helsinki Parsed Corpus of Old English Prose. University of York. Distributed through the Oxford Text Archive. http://www-users.york.ac.uk/~lang22/YcoeHome1.htm

YCOEP: Pintzuk, Susan & Leendert Plug. 2002. The York-Helsinki Parsed Corpus of Old English Poetry. Department of Linguistics, University of York. Distributed through the Oxford Text Archive, first edition, (http://www-users.york.ac.uk/~lang18/pcorpus.html).

PPCME2: Kroch, Anthony & Ann Taylor. 2000. The Penn-Helsinki Parsed Corpus of Middle English. Department of Linguistics, University of Pennsylvania. CD-ROM, second edition, release 4 (http://www.ling.upenn.edu/ppche/ppche-release-2016/PPCME2-RELEASE-4).

PCEEC: Taylor, Ann, Arja Nurmi, Anthony Warner, Susan Pintzuk, & Terttu Nevalainen. 2006. York-Helsinki Parsed Corpus of Early English Correspondence. Compiled by the CEEC Project Team. York: University of York and Helsinki: University of Helsinki. Distributed through the Oxford Text Archive.

PPCEME: Kroch, Anthony, Beatrice Santorini, & Lauren Delfs. 2004. The Penn-Helsinki Parsed Corpus of Early Modern English. Department of Linguistics, University of Pennsylvania. CD-ROM, first edition, release 3 (http://www.ling.upenn.edu/ppche/ppche-release-2016/PPCEME-RELEASE-3)

PPCMBE2: Kroch, Anthony, Beatrice Santorini, & Ariel Diertani. 2016. The Penn Parsed Corpus of Modern British English. Department of Linguistics, University of Pennsylvania. CD-ROM, second edition, release 1 (http://www.ling.upenn.edu/ppche/ppche-release-2016/PPCMBE2-RELEASE-1).

Appendix II. Search and coding queries

The data include all subjects of finite clauses which contain a conjunction or a pos-tag CONJP.

The basic set of searches to generate the data in the paper for the PPHE corpora (YCOE, PPCME, PCEEC, PPCEME, PPCMBE) are the same, modulo a different label for subjects. In the set of queries below, a definition is used for "subject" which will work for both the YCOE and later corpora. The searches for the Penn Treebank corpora, which differ rather more from the PPHE corpora, are given in section B.

A. Query files for the PPHE corpora

Note that these queries were run using CorpusSearch version 2.21. Using a later version of CorpusSearch may require a different use of the general wildcard * and the digit wildcard #.

The label ***subject*** used in the queries is a definition, and should be replaced with the appropriate label either by hand or by using the following definition file:

```
subject: NP-NOM|NP-NOM-RSP*|NP-NOM-x*|NP-SBJ*
```

A.1 Extract all coordinated subjects in finite clauses

query file: cs.q
input files: all corpus files
output file: cs.out

```
ignore_nodes: null
nodes_only: t
remove_nodes: t
node: IP-MAT*|IP-SUB*
query: (IP-MAT*|IP-SUB* iDoms ***subject***)
       AND (***subject*** iDomsMod NP !\*con*)
       AND (***subject*** iDoms CONJP|CONJ|NEG+CONJ)
```

The second line of the query is to eliminate a few cases of empty 1st conjuncts.

A.2 Separate split and non-split subjects

The `print_complement:t` command splits the input file into a file of hits that match the query (.out) and a file that doesn't (.cmp). In this case, the non-matching file contains the non-split coordinations.

query file: cs-split.q
input file: cs.out (output file from A.1)
output files: cs-split.out, cs-split.cmp

```
print_complement: t
ignore_nodes: null
node: IP-MAT*|IP-SUB*
query: (IP-MAT*|IP-SUB* iDoms ***subject***)
  AND (***subject*** iDoms CONJP)
```

```
  AND (CONJP iDoms \*ICH*)
  AND (IP-MAT*|IP-SUB* iDoms CONJP-#)
  AND (\*ICH* sameIndex CONJP-#)
```

cs-split.out contains all the split coordinated subjects (column 2 of Table 1)
cs-split.cmp contains all the non-split coordinated subjects (column 1 of Table 1)

Rename cs-split.cmp to cs-nonsplit.out for clarity

A.3 Code split coordinated subjects as split, and for the position of the 2nd conjunct (final/non-final), as illustrated in (11) for Table 3.

coding query file: c2-position.c
input file: cs-split.out (output file from A.2)
output file: c2-position.cod

```
node: $ROOT
ignore_nodes: COMMENT|CODE|ID|LB|'|\"|,|E_S|.|/
coding_query:

/* code all tokens as split */

1: {
   split: ELSE
   }

/* split: final vs non-final */

2: {
    final: ($ROOT iDoms subject)
           AND (subject iDoms CONJP)
           AND (CONJP iDoms \*ICH*)
           AND (\*ICH* sameIndex CONJP-#)
           AND ($ROOT iDomsLast CONJP-#)
    non.final: ($ROOT iDoms subject)
               AND (subject iDoms CONJP)
               AND (CONJP iDoms \*ICH*)
               AND ($ROOT iDoms CONJP-#)
               AND (\*ICH* sameIndex CONJP-#)
               AND (CONJP-# precedes *)
    }
```

A.4 Code non-split coordinations as non.split and for type (needed for coding length; cf. A.5 below)

Because of the way certain non-split subjects are parsed, it is necessary to know specific information about the type of coordination in order to count the length of the conjuncts. The easiest way to do this is to code for the different types; all others being coded as '/' (i.e. NA=not applicable).

The following types need special treatment:

(18) coordinations with multiple conjuncts (more than two): coded `mult.conjp`
(19) coordinations with shared modifiers (labelled NX in the corpus): coded `conj.x`
(20) word-level coordinations, referred to as "flat", which lack a CONJP: coded `flat`

coding query file: special-nonsplit.c
input file: cs-nonsplit.out (output file from (A.2))
output file: special-nonsplit.cod

```
node: $ROOT
ignore_nodes: COMMENT|CODE|ID|LB|'|\"|,|E_S|.|/
coding_query:

/* code all tokens as non.split */
1: {
    non.split: ELSE
    }

/* non-split: special types */
2: {
 mult.conjp: ($ROOT idoms subject)
    AND (subject idoms [1]CONJP)
    AND (subject idoms [2]CONJP)
 conj.x: ($ROOT idoms subject)
    AND (subject idoms CONJP)
    AND (CONJP iDoms NX*)
 flat: ($ROOT idoms subject)
    AND (subject idoms CONJ|NEG+CONJ)
 /: ELSE
    }
```

A.5 Code for length of conjuncts

Coding for length is a rather complicated process, as outlined in the paper. The three numbers required are the length of the 1st conjunct (L(C1)), the length of the 2nd conjunct (L(C2)), and the length of the whole subject (L(subject)). If any two of these numbers can be generated automatically by CorpusSearch, the other can be calculated in a spreadsheet. In a few cases, counting has to be done (partly) by hand as it is not possible to generate more than one measurement automatically.

Note the following:
(21) if any length is measured by CorpusSearch as > 30, it is set to 30.
(22) subjects (split or unsplit) containing clauses are coded `clause`, and not included in the length calculations, as noted in the paper

The table below summarizes schematically how lengths can be measured for various types of subjects. In Tables 4-7 in the paper, subjects with shared modification (`conj.x`) or containing clauses (`clause`) are omitted.

type of subject/split	L(C1)	L(C2)	L(subject)
subject (split or unsplit) containing a clause in any conjunct	coded 'clause'	coded 'clause'	coded 'clause'

split subject	measured by CS	measured by CS	L(C1)+L(C2)
unsplit subject with 2 conjuncts	measured by CS	L(subj) – L(C1)	measured by CS
unsplit subject with more than 1 conjunct `mult.conjp`	measured by CS	L(subj) – L(C1)	measured by CS
unsplit subject containing NX `conj.x`	manually counted	L(subj) – L(C1)	measured by CS
unsplit flat subjects `flat`	necessarily 1 word	L(subj) – L(C1)	measured by CS

A.5.1 Code subjects for length of 1st conjunct (C1) and, where possible, length of 2nd conjunct (C2)

Lengths to be calculated in a spreadsheet are coded `calculate`.

Some errors in the parsing of coordinations in the corpora are detected by the coding below and manually removed from the spreadsheet. A small number of errors, however, are simply counted wrongly. Given the amount of data, this will not appreciably affect the results reported here and we have not corrected them. Errors of this type are best corrected in the parsed files; alternatively they can be corrected in post-processing.

At this stage, split and non-split coordinations are coded for type and so are processed together.

coding query file: length-1.c
input files: c2-position.cod special-nonsplit.cod (output coded files from A.3, A.4)
output file: length-1.cod

```
node: $ROOT
ignore_nodes: COMMENT|CODE|ID|LB|'|\"|,|E_S|.|/
coding_query:

/* length of 1st conjunct in words */
3: {
 exclude: (CODING col 2 conj.x) /* exclude shared modifier (NX) type */
 \1: (CODING col 2 flat) /* assume length 1 for C1 of flat coordinations */
 clause: ($ROOT idoms subject )
     AND (subject idoms NP|NP-SBJ|NP-NOM)
     AND (NP|NP-SBJ|NP-NOM doms RMV*)
 \1: ($ROOT idoms subject)
     AND (subject idoms NP|NP-SBJ|NP-NOM)
     AND (NP|NP-SBJ|NP-NOM domsWords 1)
 \2: ($ROOT idoms subject )
     AND (subject idoms NP|NP-SBJ|NP-NOM)
     AND (NP|NP-SBJ|NP-NOM domsWords 2)

[coding for lengths 3-28 as above]

 \29: ($ROOT idoms subject)
     AND (subject idoms NP|NP-SBJ|NP-NOM)
     AND (NP|NP-SBJ|NP-NOM domsWords 29)
 \30: ($ROOT idoms subject)
```

```
     AND (subject idoms NP|NP-SBJ|NP-NOM)
     AND (NP|NP-SBJ|NP-NOM domsWords> 29)
  \1: ELSE /* leftovers are badly parsed flat split (and a few errors) */
    }

/* length of 2nd conjunct in words */
4: {
 calculate: (CODING col 1 non.split) /* C2 calculated for nonsplit type */
 clause: ($ROOT idoms subject)
     AND (subject idoms CONJP)
     AND (CONJP idoms \*ICH*)
     AND ($ROOT idoms CONJP-#)
     AND (\*ICH* sameIndex CONJP-#)
     AND (CONJP-# doms RMV*)
 \1: ($ROOT idoms subject)
     AND (subject idoms CONJP)
     AND (CONJP idoms \*ICH*)
     AND ($ROOT idoms CONJP-#)
     AND (\*ICH* sameIndex CONJP-#)
     AND (CONJP-# domsWords 1)
 \2: ($ROOT idoms subject)
     AND (subject idoms CONJP)
     AND (CONJP idoms \*ICH*)
     AND ($ROOT idoms CONJP-#)
     AND (\*ICH* sameIndex CONJP-#)
     AND (CONJP-# domsWords 2)

[coding for lengths 3-28 as above]

 \29: ($ROOT idoms subject)
     AND (subject idoms CONJP)
     AND (CONJP idoms \*ICH*)
     AND ($ROOT idoms CONJP-#)
     AND (\*ICH* sameIndex CONJP-#)
     AND (CONJP-# domsWords 29)
 \30: ($ROOT idoms subject)
     AND (subject idoms CONJP)
     AND (CONJP idoms \*ICH*)
     AND ($ROOT idoms CONJP-#)
     AND (\*ICH* sameIndex CONJP-#)
     AND (CONJP-# domsWords> 29)

/* no leftovers */
}
```

A.5.2 Code for length of whole subject, where possible

coding query file: length-2.c
input file: length-1.cod (coded file from (A.5.1))
output file: length-2.cod

```
node: $ROOT
ignore_nodes: COMMENT|CODE|ID|LB|'|\"|,|E_S|.|/
coding_query:

/* total length */
5: {
 clause: (CODING col 3 clause) /* for split subject */
 clause: (CODING col 4 clause) /* for split subject */
```

```
 clause: ($ROOT idoms subject)
     AND (subject doms RMV*)
 calculate: (CODING col 1 split) /* length calculated for split subjects */
 \3: ($ROOT idoms subject)
     AND (subject domsWords 3)

[coding for lengths 4-28 as above]

 \29: ($ROOT idoms subject)
     AND (subject domsWords 29)
 \30: ($ROOT idoms subject)
     AND (subject domsWords> 29)
 error: ELSE /* these really are mostly errors */
  }
```

B. Queries for Penn Treebank corpora (Brown, CallHome, Switchboard, Wall Street Journal)

The following queries will retrieve split and non-split coordinations, respectively, from the Penn Corpora. Note that in order to use CorpusSearch on the Penn corpora, the format must be altered slightly, as detailed here: http://corpussearch.sourceforge.net/CS-manual/YourCorpus.html

B.1. Split subject coordinations

```
node: $ROOT
query: (NAC-# iDomsFirst CC)
AND (CC iDoms and|or|nor)
AND (CC|CONJP hasSister NP*)
AND (NP-SBJ* iDoms NAC)
AND (NAC iDomsMod -NONE- \*ICH*)
```

B.2. Non-split subject coordinations

```
node: $ROOT
query: (NP-SBJ* iDomsMod NP* CC)
AND (CC iDoms and|or|nor)
```