

On the Effectiveness of Annotated Corpora: Review of the Brooklyn Corpus

YANAGI Tomohiro*
OHKADO Masayuki**

1. Introduction

In the 1960s the first computer-readable electric corpus, later to be known as the Brown Corpus, was compiled at Brown University in the United States. The corpus consisted of about one million words, containing 500 text samples taken from publications in 1961. At this phase, the Brown Corpus was a 'raw' corpus comprising plain texts, and then the frequency lists obtained from the corpus would inevitably have certain shortcomings. For this point Kennedy (1998: 209) states:

the computer could count and sort word forms, but in doing so it buried or distorted some important facts about the language. Specifically, [. . .] unless they could be lemmatized, variant inflected forms of nouns and verbs especially would be treated as entirely different word types.

He continues:

Conversely, homographs could not be distinguished. The frequency count of the number of occurrences of the word form *can* in the corpus would include the noun which refers to a metal container as well as the modal auxiliary verb.

These two problems could be overcome by annotating the corpus. Thus, soon after the completion of the Brown Corpus, the first computer corpus annotation project was undertaken: a tagset of 77 different word-class labels (see Kennedy (1998: 210-211) for the tagset) was used and the tagging was implemented automatically by computer. Such

*外国語教室 (Department of Foreign Languages)

**人文学部 (College of Humanities)

word-class tags would identify not only major parts of speech, e.g. noun, verb, preposition, and others, but also values defining subclasses, e.g. singular and plural nouns, positive, comparative and superlative adjectives, present and past verbs, and so on (cf. Leech (1997: 8)).

A second major corpus annotation project was the tagging of the British counterpart of the Brown Corpus, i.e. the LOB Corpus, which was compiled in the 1970s at Lancaster, Oslo and Bergen Universities. This time 133 basic word and punctuation tags were used. Afterwards, some kinds of annotated corpora have been developed. For example, a large-scale grammatically-tagged corpus is the British National Corpus, which contains more than 100 million words of modern English, both written and spoken; a corpus containing prosodic annotations, such as the stress, intonation, pauses and other prosodic features, is the London-Lund Corpus (cf. Garside et al. (1997)).¹⁾

The first syntactically annotated diachronic corpus, to our knowledge, is the first edition of the Penn-Helsinki Parsed Corpus of Middle English (henceforth, PPCME 1), which was compiled in 1994 (cf. Taylor and Kroch (1994)). The PPCME 1 text samples are largely based on the Middle English section of the Diachronic Part of the Helsinki Corpus of the English Texts (henceforth, the Helsinki Corpus), which has been widely used in historical studies of the English language (cf. Rissanen et al. (1993)). The PPCME 1 contains 510,000 words of text. In 1998 an improved and extended version, the PPCME 2, was released (cf. Taylor and Kroch (1998)).²⁾ The size of the samples is considerably larger: the PPCME 2 contains 1.3 million words in total. Based on the annotation scheme of the PPCME, an annotated corpus of Old English also came into being in 2000: the Brooklyn-Geneva-Amsterdam-Helsinki Parsed Corpus of Old English (henceforth, the Brooklyn Corpus), which is the target corpus of this review.³⁾

The Brooklyn corpus is an annotated version of texts from the Old English section of the Helsinki Corpus. The corpus contains more than 106,000 words from 16 Old English texts representing a range of dates of composition, authors, and genres. The Old English period is divided into the four subperiods (1=before 850, 2=850-950, 3=950-1050, and 4=1050-1150) and the texts are classified in accordance with their date of composition and the date of the manuscript. Each text is supplied in four different formats, each format as a separate file with the same name and a different extension (cf. Pintzuk (2000: 6)). A complete list of the names of the files and the relevant information is given in appendix. The four different formats are organized as follows (cf. Pintzuk (2000: 7)):

Format 1, with extension .o2/.o23/.o24/.o3/.o34. The largest version of the corpus, with a header block and three sections for each token: 1) the Old English data with syntactic annotation, 2) the gloss, and 3) the morphological annotation. This format can be searched only by search functions within word-processing programs and editors.

Format 2, with extension .oe. The Old English data with syntactic annotation only. This format can be searched by word-processing programs and by the PERL scripts supplied with the Brooklyn Corpus.

Format 3, with extension .unpacked. The Old English data unpacked, with one clause per token. This format can be searched by word-processing programs and by the PERL scripts supplied with the Brooklyn Corpus.

Format 4, with extension .psd. The Old English data in a format suitable only for searching with CorpusSearch.⁴⁾

The files in Format 1, which are our main concern, are divided into three parts: (i) syntactically annotated texts; (ii) word-for-word Present-day English gloss; and (iii) morphologically annotated texts. This is illustrated in (1)-(3).

- (1) ([+ and/CC] %nn% [vt geseah/VT] [na Agathen/NMA] [pp betwux/PP
engla/NNG werodum/NND] , 1[R [rr +anlice/RR] [vn gefretewode/VNA] R]1 ,)
(AELIVE,I,210.22)
- (2) and saw Agatha between angels throngs, singularly adorned,
- (3) and_l(and)_m(w:coco)
geseah_l(geseon)_m(m:in_n:sg_p:3_t:pa_w:vb)_c(2)
Agathen_l(agathes)_m(c:ac_n:sg_w:nona)
betwux_l(betwux)_m(w:pp)
engla_l(engel)_m(c:gn_n:pl_w:nono)
werodum_l(werod)_m(c:dt_n:pl_w:nono)
+anlice_l(+anlice)_m(w:av)
gefretewode_l(gefr+atwan)_m(c:ac_m:pt_n:sg_t:pa_w:vb)

The last part (3) contains the following information (cf. Pintzuk (2000: 8)):

Field 1: the Old English word as it appears in the text

Field 2 (beginning with ‘_l’): the Old English lemma corresponding to the word

Field 3 (beginning with ‘_m’): the morphological annotation

Field 4 (beginning with ‘_p’): for verbs with (separable) particles only, the particle

Field 5 (beginning with ‘_c’): for lexical verbs only, the complementation code

Part one (1) is an annotated Old English text. Capital letters following each word with a slash indicate the part-of-speech and inflectional information of the word. For example, “VT,” which is preceded by *geseah* ‘saw,’ is for a tensed verb; “NNG” and “NND” for genitive and dative noun, respectively (see Pintzuk (2000: section II)). Small letters next to a left bracket provide syntactic information of a phrase: e.g., “na” is for an argument in accusative and “vn” for a past participle. Clause-level syntactic tags are also added: in (1), the parentheses without a label are for the normal declarative main clause and the brackets with the label “R” for the adjunct participle/adjectival clause (see Pintzuk (2000: section III)). What is more important is that an empty category is employed. The symbol “%nn%” in (1) indicates that a nominative argument is omitted or missing. Combining capital and small letters and/or lexical words, various kinds of effective searches can be conducted. Some instances are given in section 2.

Part two (2) is word-for-word Present-day English gloss, which may be helpful for laypersons of Old English texts. It has some shortcomings, though (see section 3).

Part three (3) contains more detailed morphological information concerning word class, case, gender, person, number, mood, and tense (see Pintzuk (2000: section IV for their exclusive lists)). It also includes the lemma corresponding to a word of the sentence in the first part. This is suitable for creating a frequency word list.

The following section shows some advantages of the annotated corpus, in comparison with the non-annotated version (e.g. the Helsinki Corpus). In section 2.1 we are concerned with syntactic annotations, and turn to morphological ones in section 2.2.

2. Advantages of Annotated Corpus

Kennedy (1987: 214) suggests that ‘[t]agging enables automatic retrieval for the analysis of particular uses of a word form. [. . .] With a tagged corpus, all uses of the word *move* as a verb (or as a noun) can be retrieved.’ This is, of course, true of the research using the Brooklyn Corpus. We can automatically retrieve particular words or word strings, by using (i) the syntactic and/or (ii) the morphological tags of the corpus. Sample searches utilizing these annotations are given in the following subsections.

2.1. Syntactic Annotation

This subsection provides three sample searches which are based on the syntactic annotations of the Brooklyn Corpus. The first case is the research done by Kamiya (2001), searching for separable and inseparable particles.

The Brooklyn Corpus employs the syntactic tag, “[pa.../PA],” to refer to one of the following particles:

- (4) adun(e), æfter, aweg, (of)dune, fore, forþ, fram, geond, in, mid, niþer, of, ofer, ongean, on, onweg, to, þurh, under, up, ut, wiþ, wiþer, ymb(e)

The uses of these particles are easily obtainable by searching the corpus for the particles with the syntactic tag specified as “[pa.../PA].” The following examples are cited from Kamiya (2001: 48): ⁵⁾

- (5) a. (. . . 2[B [cs for/PP +tan/PDI] [co +te/CO] [nn +tu/PEN] [na hit/PEA] [~ne/NE] [at miht/MT] [pp mid/PP +te/PED] [pa aweg/PA] [vv l+adan/VV] B]2 .) (AELIVE,I,212.46)
 [. . . for þan þe þu hit ne miht mid þe aweg lædan (ÆLS I 212.49)
 ‘ . . . because thou canst not take it away with thee’
 b. ([+ and/CC] [rt sona/RT] [rt +ta/RT] [pa inn/PA] [vt eode/VT] [nn an/PIN +t+as/PDG cyninges/NNG +teгна/NNG 1[L %wh% [co +te/CO] %nn% [na his/PO +almyssan/NNA] [vt bewiste/VT] L]1] ,) (AELIVE,III,130.87)
 [and sona þa inn eode an þæs cyninges þegna þe his ælmyssan bewiste (ÆLS III 130.90)
 ‘And anon there came in one of the king’s thegns who had charge of his alms’

There are some cases where particles are included within prepositional phrases. In the PPCME 1, particles of this type are regarded as specifiers of the prepositional phrases, and their ‘particle’ tags are eliminated. Thus, such particles cannot be retrieved by means of the tag.

In the Brooklyn Corpus, too, some particles appear within prepositional phrases, “[pp...],” as in (6).

- (6) ([pp +after/PP +disum/PDD gebede/NND] [vt b+arst/VT] [pp ut/PA of/PP heofonum/NND] [nn swy+de/RR f+arlic/JJN fyr/NNN] ,) (AELIVE,I,316.141)

[Æfter ðisum gebede bærst ut of heofonum swyðe færlíc fyr (ÆLS I 316.141)
 ‘After this prayer, a very sudden fire burst out from Heaven’]

Since the tag “/PA,” is still added to the particle, which is contrasted to the PPCME 1, examples like (6) can also be retrieved without problem. This is one of the improvements of the Brooklyn Corpus, as Kamiya (2001: 248) notes.

Moreover, it is easy to search for those uses of the particles placed before verbs, since such particles are followed by the underscored version, “_PA”:

- (7) (. . . 2[B [co +t+at/CO] [nn hi/PEN] [na +t+at/PDA Godes/NMG m+aden/NNA]
 [pp mid/PP heora/PO galdrum/NND] [pa ofer_PA] [vt swy+ddon/VT] B]2 . .)
 (AELIVE,I,216.103)
 [. . . þæt hi þæt godes mæden mid heora galdrum oferswyðdon
 (ÆLS I 216.105)
 ‘ . . . that they with their enchantments might overpower the virgin of God.’]

Kamiya makes use of the tag “_PA,” to produce a concordance of the particles (see Kamiya (2001: Appendix 2)).

Let us next turn to the second case indicating advantages of the annotated corpus, i.e. homographs. Homographs are the words that morphologically have the same form but have different meanings or functions. A typical example of homographs is *þa*. This is assigned in the Brooklyn Corpus the three different tags: “PDN”/“PDA,” “RT” and “CS.” Each example is given in (8)-(10), with the relevant parts in bold.

- (8) ([+ and/CC] [nn ealle/JJN +**da**/PDN godas/NNN] [rr grundlunga/RR] [vt suncon/VT] [pp into/PP +t+are/PDD eor+dan/NND] ,) (AELIVE,I,316.141)
 [and ealle ða godas grundlunga suncon into þære eorðan (ÆLS I 316. 143)
 ‘and all the gods sunk utterly into the earth’]
- (9) ([+ and/CC] [rt +**da**/RT] [vt awoc/VT] [nn Lucia/NMN] .) (AELIVE,I,212.33)
 [and ða awoc lucia (ÆLS I 212.33)
 ‘and then Lucy awoke’]
- (10) ([+ Ac/CC] 1[B [cs +**da**_RT +**ta**/CS] [nd him/PED] [nn~ naht/NNN~] [~ ne/NE] [vt speow/VT] B]1 , [rt +**ta**/RT] [vt het/VT] [nn he/PEN] 2[U [vv spannan/VV] [na oxan/NNA] [pa to/PA] U]2 ,) (AELIVE,I,216.106)
 [ac ðaþa him naht ne speow þa het he spannan oxan to (ÆLS I 216.106)

‘But when they sped not at all, he commanded oxen to be harnessed to her’]

If we search the file “coaelive.o3” for *ba* without using a tag, we will find 222 tokens. This result includes all the three types of *ba*, mentioned above. Regardless of which of the three words is the target word, the irrelevant instances have to be manually put away for the ‘correct’ count of the target word. By contrast, the use of a tag makes it easier to collect data needed. When the same file is searched by specifying the tag of *ba* as “PDN” and “PDA,” 78 instances of the use of *ba* as a determiner will be obtained more quickly.

The last case showing advantages of the annotated corpus resides in a search for a particular clause type of sentence: imperative, exclamation, embedded main clauses, infinitival clauses, and other clause types. For example, we will easily obtain all the tokens of imperatives in the file “coaelive.o3,” by searching for the tags “(V... V)” and “[V ... V].” 26 instances are included in the file. Here are three examples:

- (11) (V [vt geneal+ac/VI] [rt nu/RT] [rt +arest/RT] V) (AELIVE,I,308.33)
 [genealæc nu ærest (ÆLS I 308.34)
 ‘Approach now first [of all]’]
- (12) (1(V [o God/NMN] [vt gemiltsa/VI] [nd urum/POD sawlum/NND] . V)1)
 (AELIVE,III,136.158)
 [God gemiltsa urum sawlum. (ÆLS III 136.161)
 ‘God, have mercy on our souls.’]
- (13) (V [+ and/CC] [na +ta/PDA +ahta/NNA] [vt ateoh/VI] 1[V [vt loca/VI] 2[H
 [wh-1 hu/WR] %co% %ne% %rr-1% [nd +te/PED] [vt licige/VT] , H]2 [o
 min/PON leofa/JJN dohtor/NNN] V]1 . V) (AELIVE,I,212.44)
 [and þa æhta ateoh loca hu þe licige min leofa dohtor (ÆLS I 212.44)
 ‘and then dispose of the property lo! how thou mayst please, my dear
 daughter’]

(11) is an example of main clause imperatives, and (12) an example of imperatives embedded in direct speech. In the latter case, the same number is added both before the left parenthesis and after the right one, just like “1(V ... V)1.” The imperative of (12) is extracted from the following main clause:

- (14) ([+ and/CC] %nn% [rr +tus/RR] [vt clypode/VT] [pp on/PP his/PO fylle/NND]
 {V1} ,) (AELIVE,III,136.158)

[and þus clypode on his fylle
'and thus cried in his fall']

(*ÆLS* III 136.161)

The 'base' position of (12) is marked with "{V1}" in (14). Since embedded imperatives like (12) stand as independent tokens, the total number of imperatives in a text sample is easy to count by making use of the tag. When an embedded imperative is not direct speech, it is enclosed in square brackets "[V ... V]," as in (13). Thus, imperatives are distinguished in three ways: main clause imperatives, and imperatives embedded in direct speech and not in direct speech. Such a distinction by syntactic tag may be useful for research of some differences in behavior of imperatives with respect to clause types.

2.2. Morphological Annotation

In the morphological annotations, lemmatization plays an important role in comparing the frequencies of synonyms in a single text or across texts. Let us take the comparison of the frequencies of the three verbs of saying, *sprecan*, *cweðan* and *secgan*, as an example.

As noted in section 1, unless they are lemmatized, inflected variants of these verbs are counted as completely different words and the correct frequency of *sprecan*, *cweðan* or *secgan* cannot be automatically obtained. Moreover, since in Old English the thorn and eth were interchangeable, variant forms of these verbs must be taken into consideration in their searches, if they are not lemmatized.

The results obtained from searching the files "coboeth.o2," "cobede.o2" and "coaelive.o3" for the lemmas *sprecan*, *cweðan* and *secgan* are given below:

(15)	<i>sprecan</i>	<i>cweðan</i>	<i>secgan</i>	total
coboeth.o2	20 (8.1%)	167 (67.6%)	60 (24.3%)	247
cobede.o2	5 (7.6%)	33 (50.7%)	27 (41.5%)	65
coaelive.o3	4 (7.7%)	33 (63.4%)	15 (28.8%)	52

This table suggests that the proportion of *sprecan* to the other verbs in parentheses is relatively constant in every work, and that the proportions of the other verbs show unsteadiness.⁶⁾

This section concludes with the following quote:

As long as the automatic tagger is an accurate one or there has been manual correction, a grammatically tagged corpus can quickly and easily be searched for

only those uses of the word *since* where it functions as a preposition.

(Kennedy (1987: 214))

3. Some Problems

Useful as the Brooklyn Corpus is, as indicated in the preceding section, there are some problems which its users should be aware of in conducting their searches.

3.1 Syntactic Annotation

According to the manual (Pintzuk (2000:9)) “[t]he annotation schemes are general-purpose ones that are as theory-neutral as possible, while still incorporating the insights of modern linguistic theory.” One of the most theory-oriented parts of the corpus is the introduction of annotations for empty categories, which is potentially very useful. Thanks to this practice, we can, for instance, search for subjectless coordinate clauses as illustrated in (16), with relative ease, through the “%nn%” tag, which stands for empty NPs in the nominative case.

- (16) ([+and/CC] %nn% [pp mid/PP manegum/JJD +deowracum/NND] [na +t+at/PDA mancynn/NNA] [vt geegsode/VT] , 1[B [co +t+at/CO] [nn hi/PEN] [na heora/PO lac/NNA] [vt geoffrodon/VT] [nd +tam/PDD leasum/JJD godum/NND] [pp mid/PP him/PED]B]1 .) (AELIVE,I,308.8)
 [and mid manegum ðeowracum þæt mancynn geegsode þæt hi heora lac geoffrodon þam leasum godum mid him (ÆLS I 308.10)
 ‘and, with many threats, frightened the people so that they offered their sacrifices to the false gods with him’]

This practice, however, is not consistently carried out for NPs that are not in the nominative case. In the following example, empty accusative objects are tagged as “%na%” in line with the empty nominative NPs.

- (17) ([+ and/CC] %nn% [vt het/VT] 1[U [vv geniman/VV] [na +tone/PDA dry/NNA 2[L %wh% [co +te/CO] %nn% [rl +t+ar/RL] [vt gelyfde/VT] [pp on/PP God/NMA] L]2 U]1 , 3[U [+ and/CC] %na% [vv l+adan/VV] [pp of/PP +d+are/PDD byrig/NND] U]3 4[U [+ and/CC] %na% [vv beheaf dian/VV] [rt sona/RT] U]4 .) (AELIVE,I,312.82)

[and het geniman þone dry þe þær gelyfde on god and lædan of ðære byrig and
beheafðian sona (ÆLS I 312.83)

‘and bade seize on the sorcerer, who had there believed in God, and bring him
out of the city, and instantly behead him’]

In contrast, no empty accusative NP object is assumed in the following example, where the finite transitive verb of the conjoined clause, *tugon* ‘tugged’ should be analyzed to take one.

- (18) ([rt +ta/RT] [vt cnitton/VT] [nn hi/PEN] [na rapas/NNA] [pp mid/PP
re+dum/JJD anginne/NND] [d hire/PED] [pp to/PP handum/NND and/CC
fotum/NND] ,) (AELIVE,I,216.100)

([+ and/CC] [nn fela/PIN] [rr samod/RR] [vt tugon/VT] ,) (AELIVE, I,
216.100)

[þa cnitton hi rapas mid reðum an-ginne hire to handum and fotum and fela
samod tugon (ÆLS I 216.100)

‘Then they fastened ropes, in their cruel attempt, to her hands and feet, and
many tugged at once’]

3.2 Morphological Annotation

It is well-known that the inflectional system of Old English is “defective” in the sense that more than one function can be expressed by one form. For instance, irrespective of the mood (indicative or subjunctive), the first and the third singular past form of weak verbs is *-de*. Such ambiguity does not seem to be taken into consideration in the corpus, so that even in ambiguous cases words are uniquely tagged. In the following example, *aspende* ‘spent’ is tagged as subjunctive (m:su), although it can also be interpreted as indicative.

- (19) ([rt Nu/RT] [nn ic/PEN] [at wylle/MT] [na me/PEA sylfe/PFA] [nd him/PED]
[rr so+dlice/RR] [vv geoffrian/VV] , 1[B [cs for+dan/CS] [nn ic/PEN] [rt
leng/RT] [vt~ n+abbe/VT~] 2[H [wh-1 hw+at/PIA] %co% %na-1% [nn ic/PEN]
[pp on/PP his/PO lacum/NND] [vt aspende/VT] H]2 B]1 .) (AELIVE,I,214.66)

now I will me self him truly offer, because I longer not-have anyone I on his
plays spend.

Nu_l(nu)_m(w:avti)
 ic_l(ic)_m(c:nm_n:sg_p:1_w:prpe)
 wylle_l(willan)_m(m:in_n:sg_p:1_t:pr_w:vbmo)
 me_l(ic)_m(c:ac_n:sg_p:1_w:prpe)
 sylfe_l(self)_m(c:ac_g:fm_n:sg_w:prrf)
 him_l(ic)_m(c:dt_n:sg_p:3_w:prpe)
 so+dlice_l(so+tlice)_m(w:av)
 geoffrian_l(geoffrian)_m(m:ib_w:vb)_c(3)
 for+dan_l(for+t+am)_m(w:cosu)
 ic_l(ic)_m(c:nm_n:sg_p:1_w:prpe)
 leng_l(leng)_m(w:avti)
 n+abbe_l(habban)_m(m:in_n:sg_p:1_t:pr_w:vbng)_c(6)
 hw+at_l(hwa)_m(c:ac_g:nt_n:sg_w:prid)
 ic_l(ic)_m(c:nm_n:sg_p:1_w:prpe)
 on_l(on)_m(w:pp)
 his_l(his)_m(w:prpo)
 lacum_l(lac)_m(c:dt_n:pl_w:nono)
 aspende_l(aspendan)_m(m:su_n:sg_t:pr_w:vb)_c(2)

[nu ic wylle me sylfe him soðlice geoffrian forðan ic leng næbbe hwæt ic on his lacum aspende. (ÆLS I 214.66)]

‘Now I desire verily to offer to Him myself, because for some time I have had nothing to spend in His service.’]

For this reason, in cases where potentially ambiguous forms are involved, searches should not be conducted blindly depending on the morphological annotation.

3.3 Discrepancies between Syntactic and Morphological Annotations

There are a number of cases in which syntactic tags are contradictory to morphological tags. In (20) below, for example, the nonfinite verb is morphologically coded as “c(4),” which indicates that the verb takes an NP and a finite clause as complements. In the text, however, it is used as a ditransitive verb, taking the dative personal pronoun *eow* ‘you’ and the accusative demonstrative pronoun *bæt* ‘that’ modified by a relative clause as complements.

- (20) ([rt Nu/RT] [at wille/MT] [nn we/PEN] [nd eow/PED] [vv secgan/VV] [na +t+at/PDA 1[L %wh% %co% %nn% [jn so+d/JJN] [vt is/BT] [pp be/PP +dam/PDD] L]1] , 2[B [co +t+at/CO] [nn heora/PO gedwyld/NNN] [~ ne/NE] [vt derige/VT] [rr digellice/RR] [nd +anigum/PID] B]2 .) (AELIVE,I,306.3)
 now will we you say that true is by that, that their wandering not damage secretly any.

Nu_l(nu)_m(w:avti)
 wille_l(willan)_m(m:in_n:sg_p:1_t:pr_w:vbmo)
 we_l(ic)_m(c:nm_n:pl_p:1_w:prpe)
 eow_l(ic)_m(c:dt_n:pl_p:2_w:prpe)
 secgan_l(secgan)_m(m:ib_w:vb)_c(4)
 +t+at_l(se)_m(c:nm_g:nt_n:sg_w:prrl)
 so+d_l(so+t)_m(c:nm_g:nt_n:sg_w:ajpr)
 is_l(beon)_m(m:in_n:sg_p:3_t:pr_w:vbcp)
 be_l(be)_m(w:pp)
 +dam_l(se)_m(c:dt_n:sg_w:prde)
 +t+at_l(+t+at)_m(w:cosu)
 heora_l(hiera)_m(w:prpo)
 gedwyld_l(gedwild)_m(c:nm_n:sg_w:nono)
 ne_l(ne)_m(w:avng)
 derige_l(derian)_m(m:su_n:sg_t:pr_w:vb)_c(2)
 digellice_l(diegollice)_m(w:av)
 +anigum_l(+anig)_m(c:dt_n:sg_w:prid)

[Nu wille we eow secgan þæt soð is be ðam þæt heora gedwyld ne derige digellice ænigum. (ÆLS I 306.3)

‘Now will we tell you that which is true about him, that their error may not secretly harm any one.’]

A list of complementation codes is given in (21).

- (21) Complementation codes:
 _c(1) intransitive

_c(2)	monotransitive
_c(3)	ditransitive
_c(4)	NP + finite clause
_c(5)	NP + non-finite clause
_c(6)	finite clause
_c(7)	non-finite clause
_c(8)	complex transitive (small clause object)
_c(9)	impersonal

In the following example the pronoun *hire* 'her' is tagged as a dative NP of possession, so that the verb should be regarded as an intransitive. However, it is analyzed as a monotransitive verb, annotated with "c(2)."

- (22) ([nn heo/PEN] [vt wear+d/ET] [rt +ta/RT] [vn gewundod/VN] , 1[B [co +t+at/CO] [d hire/PED] [vt wand/VT] [nn se/PDN inno+d/NNN] [pa ut/PA] B]1 ,) (AELIVE,I,216.127)

she became then wounded, that her wound that inside out,

heo_l(ic)_m(c:nm_g:fm_n:sg_p:3_w:prpe)
wear+d_l(wear+tan)_m(m:in_n:sg_p:3_t:pa_w:vbax)
+ta_l(+ta)_m(w:avti)
gewundod_l(gewundian)_m(m:pt_t:pa_w:vb)_c(2)
+t+at_l(+t+at)_m(w:cosu)
hire_l(ic)_m(c:dt_g:fm_n:sg_p:3_w:prpe)
wand_l(windan)_m(m:in_n:sg_p:3_t:pa_w:vb)_c(2)
se_l(se)_m(c:nm_g:ma_n:sg_w:prde)
inno+d_l(inno+t)_m(c:nm_n:sg_w:nono)
ut_l(ut)_m(w:pa)

[heo wearð þa gewundod þæt hire wand se innoð ut (ÆLS I 216.127)
'Then was she wounded, so that her bowels fell out']

In the following example the nonfinite verb *gehyran* 'hear' takes an accusative NP accompanied by a present participle. In the morphological annotation, however, it is

analyzed as a verb taking a nonfinite clause, i. e. “c(7).”

- (23) (1(D 2[B [cs +teah/CS] [nn +tu/PEN] [vt clypige/VT] [na tyn/JJA
+tuseþ/NNA manna/NNG] B]2 , [nn hi/PEN] [at sceolan/MT] [en ealle/JJN]
[vv gehyran/VV] [na +tone/PDA halgan/JJA gast/NNA] 3[U [rr +tus/RR] [vg
cwe+dende/VG] , {D2} U]3 D)1) (AELIVE,I,216.111)

although you speak ten thousand persons, they must all hear that holy breath
thus saying,

+teah_l(+teah)_m(w:cosu)
+tu_l(ic)_m(c:nm_n:sg_p:2_w:prpe)
clypige_l(clypian)_m(m:su_n:sg_t:pr_w:vb)_c(2)
tyn_l(tien)_m(c:ac_n:sg_w:nrcd)
+tuseþ_l(+tuseþ)_m(c:ac_n:sg_w:nrno)
manna_l(mann)_m(c:gn_n:pl_w:nono)
hi_l(ic)_m(c:nm_n:pl_p:3_w:prpe)
sceolan_l(sculan)_m(m:in_n:pl_t:pr_w:vbmo)
ealle_l(eall)_m(c:nm_n:pl_w:ajat)
gehyran_l(gehieran)_m(m:ib_w:vb)_c(7)
+tone_l(se)_m(c:ac_g:ma_n:sg_w:prde)
halgan_l(halig)_m(c:ac_n:sg_w:ajat)
gast_l(gast)_m(c:ac_n:sg_w:nono)
+tus_l(+tus)_m(w:av)
cwe+dende_l(cwe+tan)_m(m:pt_t:pr_w:vb)_c(6)

[beah þu clypige tyn þuseþ manna hi sceolan ealle gehyran þone halgan gast
þus cweðende (ÆLS I 216. 111)

‘though thou call ten thousand men, they would all hear the Holy Ghost thus
speaking:’]

3.4 Gloss

As the following citation from the manual (Pintzuk (2000: 11)) shows, the glosses of the corpus were automatically constructed so that they are often misleading, or simply “wrong.”

Part 2 of each token contains the word-for-word Modern English gloss. These glosses were constructed automatically, without regard for sense or reasonable translation, simply by using the first definition for each lemma (=base form) from Clark Hall 1960. It should be emphasized that the glosses are meant only as an aid, are not guaranteed to be accurate, and in fact frequently do not make much sense. Use them at your own risk! Better yet, do not use them at all!

For instance, the gloss ‘without’ of *leas* ‘false’ in the following example will only confuse the reader.

- (24) ([+ and/CC] %nn% [pp mid/PP manegum/JJD +deowracum/NND] [na +t+at/PDA mancynn/NNA] [vt geegsode/VT] , 1[B [co +t+at/CO] [nn hi/PEN] [na heora/PO lac/NNA] [vt geoffrodon/VT] [nd +tam/PDD leasum/JJD godum/NND] [pp mid/PP him/PED] B]1 .) (AELIVE,I,308.8)
 [and mid manegum ðeowracum þæt mancynn geegsode þæt hi heora lac geoffrodon þam leasum godum mid him (ÆLS I 308.10)
 ‘and, with many threats, frightened the people so that they offered their sacrifices to the false gods with him’]

Maybe, we should follow the editors suggestion and should not use the glosses at all.

4. Summary

In the present review we have shown that, as illustrated by the sample searches in section 2, the Brooklyn Corpus is a potentially very useful tool for syntactic studies of Old English. We believe that, combined with the Penn-Helsinki Parsed Corpus of Middle English, a syntactically annotated corpus of Middle English prose, the Brooklyn Corpus can also serve as an effective tool for diachronic studies of the first two periods of the English language (the periods of Old and Middle English). For more extended diachronic survey of English, completion of the Penn-Helsinki Parsed Corpus of Modern English, which is under construction at the University of Pennsylvania by Anthony Kroch and Beatrice Santorini, is looked forward to.⁷⁾

Notes

- 1) The Brown Corpus, the LOB Corpus and the London-Lund Corpus are contained in the second version of the ICAME Corpus Collection on CD-ROM. More information is available at <http://nora.hd.uib.no/icame/cd/>. For the British National Corpus see <http://thetis.bl.uk/>.
- 2) More information of the PPCME, first and second editions is available at <http://www.ling.upenn.edu/mideng/>. The Helsinki Corpus is contained in the second version of the ICAME Corpus Collection. See note 1 for its site.
- 3) The Brooklyn Corpus is available without fee for educational and research purposes. More information is obtainable at <http://www-users.york.ac.uk/~sp20/corpus.html>
- 4) CorpusSearch is a powerful java-based search engine developed for use with the PPCME 2. It can also be used with other corpora annotated in the Penn Treebank style, including the Brooklyn Corpus.
- 5) In this example and those following, a non-annotated version of Old English text and its Present-day English translation of Skeat (1881-1900), which are in brackets, are added for readers' convenience.
- 6) For detailed discussin of these verbs see Ogura (1981), who compares *cyðan*, *reccan* and *boddian* as well.
- 7) The York-Toronto-Helsinki Parsed Corpus of Old English, a larger corpus of Old English, with more detailed syntactic annotation is also under construction at the University of York by Ann Taylor, Anthony Warner, Susan Pintzuk, and Frank Beths. More information of this corpus is available at <http://www-users.york.ac.uk/~lang22/ycoc.htm>.

References

- Clark Hall, John R. (1960) *A Concise Anglo-Saxon Dictionary*, 4th edn, with a supplement by Herbert Dean Meritt. Cambridge University Press, Cambridge.
- Garside, Roger, Geoffrey Leech and Anthony McEnery (ed.) (1997) *Corpus Annotation: Linguistic Information from Computer Text Corpora*, Longman, London/New York.
- Kamiya, Masaaki (2001) "The Syntactic Annotation of the Brooklyn Corpus," *Journal of Toyota National College of Technology* 34, 245-254. [in Japanese].
- Kennedy, Graeme (1998) *An Introduction to Corpus Linguistics*, Longman, London/New York.
- Leech, Geoffrey (1997) "Introducing Corpus Annotation," in Garside, Leech and McEnery (1997), 1-19.
- Ogura, Michiko (1981) *The Syntactic and Semantic Rivalry of Quoth, Say and Tell in Medieval English*, KUFs Publication, Osaka.
- Pintzuk, Susan (2000) *Manual of the Brooklyn-Geneva-Amsterdam-Helsinki Parsed Corpus of Old English*.
- Rissanen, Matti, Merja Kytö, Minna Palander-Collin (eds.) (1993) *Early English in the Computer Age: Explorations through the Helsinki Corpus*, Mouton de Gruyter, Berlin.
- Skeat, Walter W. (1881-1900) *Ælfric's Lives of Saints*, EETS 76, 82, 94, 114 [repr. as two volumes in 1966].

Taylor, Ann and Anthony S. Kroch (1994) *Manual of the Penn-Helsinki Parsed Corpus of Middle English*.

Taylor, Ann and Anthony S. Kroch (1998) *Manual of the Penn-Helsinki Parsed Corpus of Middle English*, Phase II.

Appendix

coaelet3.o3

AELFRIC'S FIRST AND SECOND LETTERS TO WULSTAN. DIE HIRTENBRIEFE AELFRICS IN ALTENGLISCHER UND LATEINISCHER FASSUNG. BIBLIOTHEK DER ANGELSAECHSISCHEN PROSA, IX. ED. B. FEHR. HAMBURG: VERLAG VON HENRI GRAND, 1914. PP. 68.12 - 144.2 (IV; LWSTAN1); PP. 146.1 - 220.7 (V; LWSTAN2) (SAMPLE 1)

word count: 7960

token reference format: page.line, where line = line within text

coaelet4.o34

AELFRIC'S LETTER TO WULFSIGE. DIE HIRTENBRIEFE AELFRICS IN ALTENGLISCHER UND LATEINISCHER FASSUNG. BIBLIOTHEK DER ANGELSAECHSISCHEN PROSA, IX. ED. B. FEHR. HAMBURG: VERLAG VON HENRI GRAND, 1914. PP. 1.15 - 34.9 (I) (SAMPLE 2)

word count: 3230

token reference format: page.line, where line = line within text

coaelive.o3

AELFRIC'S LIVES OF SAINTS, BEING A SET OF SERMONS ON SAINTS' DAYS FORMERLY OBSERVED BY THE ENGLISH CHURCH, VOLS. I-II. EARLY ENGLISH TEXT SOCIETY, O.S. 76, 82, 94, 114. ED. W. W. SKEAT. LONDON, 1966 (1881-1900). PP. 210.1 - 218.152 (IX) (SAMPLE 1); PP. 306.1 - 318.184 (XIV) (SAMPLE 2); PP. 124.1 - 142.288 (XXVI) (SAMPLE 3); PP. 314.1 - 334.276 (XXXII) (SAMPLE 4)

word count: 6980

token reference format: page.line, where line = line within life

coapollo.o3

THE OLD ENGLISH 'APOLLONIUS OF TYRE'. ED. P. GOOLDEN. LONDON: OXFORD

UNIVERSITY PRESS, 1958. PP. 2.1 - 42.31

word count: 6530

token reference format: page.section.line, where line = line within section

cobede.o2

BEDE'S ECCLESIASTICAL HISTORY. THE OLD ENGLISH VERSION OF 'BEDE'S ECCLESIASTICAL HISTORY OF THE ENGLISH PEOPLE', PARTS I,1; I,2. EARLY ENGLISH TEXT SOCIETY, O.S. 95, 96. ED. T. MILLER. LONDON, 1959 (1890; 1891). PP. 104.12 - 124.24 (SAMPLE 1); PP. 252.4 - 272.18 (SAMPLE 2); PP. 418.22 - 436.19 (SAMPLE 3)

word count: 10170

token reference format: section.page.line, where line = line within page

coboeth.o2

ALFRED'S BOETHIUS. KING ALFRED'S OLD ENGLISH VERSION OF BOETHIUS DE CONSOLATIONE PHILOSOPHIAE. ED. W. J. SEDGEFIELD. OXFORD: THE CLARENDON PRESS, 1899. PP. 74.15 - 79.7 (XXXIII) (SAMPLE 1); PP. 82.18 - 89.3 (XXXIV) (SAMPLE 2); PP. 89.24 - 94.13 (SAMPLE 3); PP. 117.1 - 124.1 (XXXVIII) (SAMPLE 4); PP. 126.28 - 135.21 (XXXIX) (SAMPLE 5)

word count: 10920

token reference format: page.line, where line = line within page

cochroa2.o2

CHRONICLE MS A EARLY (O2). TWO OF THE SAXON CHRONICLES PARALLEL, VOL. I. ED. C. PLUMMER. OXFORD: THE CLARENDON PRESS, 1965 (1892). PP. 2.1 - 112.5

word count: 13460

token reference format: page.(year).line, where line = line within year

cochroa3.o3

CHRONICLE MS A LATE (O3). TWO OF THE SAXON CHRONICLES PARALLEL, VOL. I. ED. C. PLUMMER. OXFORD: THE CLARENDON PRESS, 1965 (1892). PP. 112.6 - 132.29

word count: 670

token reference format: page.(year).line, where line = line within year

cogregd3.o23

GREGORY THE GREAT, DIALOGUES (MS H). BISCHOF'S WAERFERTH VON WORCESTER UEBERSETZUNG DER DIALOGE GREGORS DES GROSSEN. BIBLIOTHEK DER ENGELSAECHSISCHEN PROSA, V. ED. H. HECHT. LEIPZIG: GEORG H. WIGAND'S VERLAG, 1900. PP. 16.6 - 25.28 (SAMPLE 1); PP. 51.32 - 61.19 (SAMPLE 2); PP. 134.22 - 143.23 (SAMPLE 3)

word count: 5170

token reference format: section.page.line, where line = line within page

cogregd4.o24

GREGORY THE GREAT, DIALOGUES (MS C). BISCHOF'S WAERFERTH VON WORCESTER UEBERSETZUNG DER DIALOGE GREGORS DES GROSSEN. BIBLIOTHEK DER ENGELSAECHSISCHEN PROSA, V. ED. H. HECHT. LEIPZIG: GEORG H. WIGAND'S VERLAG, 1900. PP. 34.23 - 43.17 (SAMPLE 1); PP. 70.34 - 79.7 (SAMPLE 2); PP. 123.17 - 133.12 (SAMPLE 3)

word count: 5100

token reference format: page.line, where line = line within page

colaw2.o2

ALFRED'S INTRODUCTION TO LAWS, LAWS (ALFRED), AND LAWS (INE). DIE GESETZE DER ANGELSACHSEN, VOL. I. ED. F. LIEBERMANN. HALLE: MAX NIEMEYER, 1903. PP. 26.17 - 46.23 (ALFRED'S INTRODUCTION TO LAWS) AND PP. 46.17 - 88.7 (LAWS (ALFRED)) (SAMPLE 1); PP. 88.25 - 122.32 (LAWS (INE)) (SAMPLE 2)

word count: $1950 + 3300 + 2670 = 7920$

token reference format: page.section.subsection, where subsection either follows Liebermann's numbering or else is the sentence number within the section.

colaw3.o3

LAWS (ELEVENTH CENTURY). DIE GESETZE DER ANGELSACHSEN, VOL. I. ED. F. LIEBERMANN. HALLE: MAX NIEMEYER, 1903. PP. 236.23 - 246.4 (V AETHELRED) PP. 246.9 - 256.27 (VI AETHELRED) PP. 278.10 - 296.5 (I CNUT) PP. 308.1 - 326.12 (II CNUT) PP. 380.1 - 385.5 (NORTHYMBRA PREOSTA LAGU)

word count: 6900

token reference format: page.section.subsection, where subsection either follows Liebermann's numbering or else is the sentence number within the section.

colaw4.o34

LAWS (LATE). DIE GESETZE DER ANGELSACHSEN, VOL. I. ED. F. LIEBERMANN. HALLE: MAX NIEMEYER, 1903. PP. 444.10 - 453.6 (RECTITUDINES) AND PP. 453.10 - 455.48 (GEREFA) (SAMPLE 1); PP. 483.1 - 484.24 (WILLIAM I) (SAMPLE 2)
word count: 2100 + 220 = 2320

token reference format: page.section.subsection, where subsection either follows Liebermann's numbering or else is the sentence number within the section.

coorosiu.o2

ALFRED'S OROSIUS. KING ALFRED'S OROSIUS, PART I. EARLY ENGLISH TEXT SOCIETY, O.S. 79. ED. H. SWEET. LONDON, 1959 (1883). PP. 58.13 - 78.30 (SAMPLE 1); PP. 102.1 - 120.18 (SAMPLE 2); PP. 228.1 - 238.14 (SAMPLE 3)
word count: 8640

token reference format: page.line, where line = line within page

cowulf3.o3

WULFSTAN'S HOMILIES (O3). THE HOMILIES OF WULFSTAN. ED. D. BETHURUM. OXFORD: THE CLARENDON PRESS, 1957. PP. 175.3 - 184.180 (VIIIc) PP. 200.3 - 210.203 (Xc) PP. 225.3 - 232.106 (XIII) PP. 267.7 - 275.202 (XX)
word count: 6950

token reference format: page.line, where line = line within homily

cowulf4.o34

WULFSTAN'S HOMILIES (O3/4). THE HOMILIES OF WULFSTAN. ED. D. BETHURUM. OXFORD: THE CLARENDON PRESS, 1957. PP. 116.3 - 118.40 (Ib) PP. 119.28 - 122.73 (II) PP. 123.4 - 127.80 (III) PP. 221.3 - 224.95 (XII) PP. 242.8 - 245.79 (XVII)

word count: 3290

token reference format: page.line, where line = line within homily

Total word count: 106,210