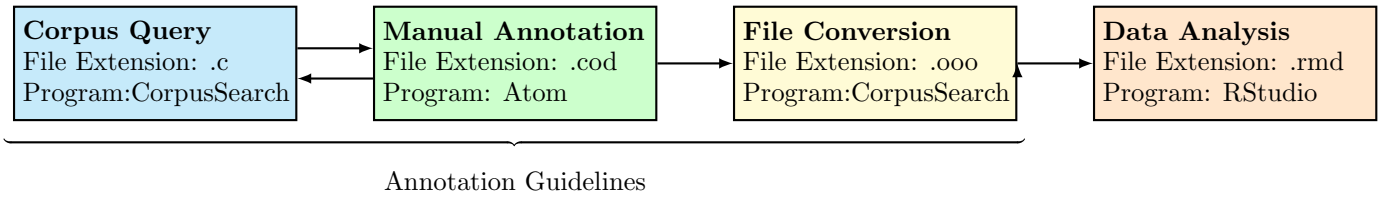
Annotation Guidelines:
The Development of Syntagmatic Redundancy in ME

Raquel Montero
STARFISH
(Dated: September 22, 2023)

These guidelines explain the annotation procedure carried out for investigating the loss of plural agreement in Quantifiers and Adjectives during the Middle English (ME) period. Two major syntactic categories were analysed: quantifiers and strong adjectives. The work is part of the STARFISH Project (*Sociolinguistic Typology and Responsive Features in Syntactic History*), funded by he European Union's Horizon 2020 research and innovation programme under grant agreement No. 851423.

I. PROCEDURE

There were four main phases in the data analysis process: Corpus Query, Manual Annotation, File Conversion and Data Analysis. Firstly, the desired NPs were searched for the in the corpus–files with extension .c–, then the output file(.cod) was manually annotated to check for inconsistencies in the original search. Based on this first exploration of the data, a second Corpus Query was run and again corrected for typos. Finally, this last document was exported as a .ooo file using the code in the file code.q and used for analysis in R. This document focuses on the first three phases, for information on how the data was processed and analysed the reader is referred to the document: `ME_Case_Study_Agreement_Stats_Graphs.rmd`



II. CORPUS QUERY

In order to subtract from the *The Penn-Helsinki Parsed Corpus of Middle English* (PPCME2) all the tokens containing a plural noun and a quantifiers or strong adjective from the corpus, the program CorpusSearch was used. A coding query was written (see CorpusSearchSyntRedMEAdjectives.c and CorpusSearchSyntRedMEQuantifiers.c), which created a string of 6 different categories for every NP found in the Corpus, as illustrated in 1:

(1) (NP (**CODING-NP c1:c2:c3:c4:c5:c6**)
(ADJ adjective)
(NS plural noun)))

A. Quantifiers

Regarding the quantifiers every NP was annotated for the following categories (z was used as an elsewhere case):

Categories	**Quantifier**	**Agreement**	**Text**	**Year**	**Region**	**Syllables**
Values	quant z	agreement no-agreement z	Chronicles Vices . . .	1150-1490	Southern EM WM Northern z	monosyllabic polysyllabic z

The category **Quantifiers** encoded weather the NP contained a plural noun(NS) and a quantifier. If that was the case the label *quant* was assigned to it, if either of the two conditions were not satisfied the label z was assigned to the string.

The category **Agreement** encoded weather the quantifier had plural morphology attached to it (agreement) or not (no agreement). Because the Penn Corpus is not lemmatised, there was no direct way of doing this. First we just run a search for all quantifiers ending in -e but this search had to be manually corrected. During the annotation process a list of all the quantifiers (and their spelling variants) was compiled and then divided depending on whether the quantifier had plural morphology or not (see the section Manual Annotation for more details). Once this was done a second search was run, including all quantifiers that were coded as agreeing:

1. First Query: agreement: (NP* iDoms Q*) AND (Q* iDoms *e)
2. Second Query: agreement: (NP* iDoms Q*) AND (Q* iDoms +alle|ale|alle|Alle|ALLE...)

The same procedure was followed for the non-agreeing quantifiers.

Regarding the categories **Text**, **Year** and **Region**, we followed the information provided in the section *Philological Information* of the PPCME2 (https://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-4/index.html). Each text was given a name, approximate date/period of composition (1150-1500), and a dialect/region(West Midlands, East Midlands, Northern, Southern).

Finally, regarding the category **syllables**, each quantifier was categorised as monosyllabic or polysyllabic. Because the corpus does not encode this type of information, first we run a search with one monosyllabic and polysyllabic quantifier (the rest where assigned a z label). After that we proceeded to the manual annotation, where we created a list of the different quantifiers, their spellings and number of syllables. A second query was run including all these quantifiers, and later double corrected in a second annotation phase:

1. First Query: monosyllabic: (NP* iDoms Q*) AND (Q* iDoms som)
2. Second Query: monosyllabic: (NP* iDoms Q*) AND (Q* iDoms som Som|ssum|sum|Sum|zome|Zome|some...)

B. Strong Adjectives

Categories	**Adjective**	**Agreement**	**Text**	**Year**	**Region**	**Syllables**
Values	postnominal prenominal conjoined multiple z	agreement no-agreement french-agree adverbial-ly e-ending z	Chronicles Vices . . .	1150-1490	Southern EM WM Northern z	monosyllabic polysyllabic z

The category **Adjective** had 5 different labels: prenominal, postnominal, conjoined, multiple and z.

In early English strong adjectives occurred in two positions: in postnominal position and in prenominal position in the absence of definitness markers (), hence the need for more categories in the case of adjectives.

- postnominal: all NPs containing a plural noun and a postnominal adjective.
- prenominal: all NPs containing a prenominal adjective and plural noun (which did not start with a vowel, as this can affect agreement in the preceding adjective) and that in addition did not have any determiners, quantifiers or possessive pronouns.

In addition to this, because each NP could only be given a string of values, NPs which contained multiple adjectives (either conjoined or stacked) were coded into a separate category and left out from the analysis:

- conjoined: if the NP contained two adjectives conjoined with AND.
- multiple: if the NP contained multiple adjectives stacked.

The category **Agreement** also had the categories agreement vs no agreement (see the above section on quantifiers to see the procedure to search for them). Additionally, there were 3 other values included:

- french-agree: there were several adjectives whose agreement was borrowed from French. For example the adjective *perdurables* has the French plural marking -s, and not the English one. These adjectives were not included in the analysis as its hard to know if they are reflections of the French grammar rather than the English one.

- adverbial-ly: adjectives ending in -ly had already drop the agreement earlier than all other adjectives, and by the early Middle English period they contain no agreement.
- e-ding: there were several adjectives whose root already ends in -e, so it is not possible to determine if the adjective shows plural agreement as the plural marking is also -e. In order to check for this we used the OE dictionary *Bosworth Toller's Anglo-Saxon Dictionary Online*(https://bosworthtoller.com/) and searched for the root of each of the different adjectives. If the adjective already included an -e in its root, it was coded as e-ending and excluded from the analysis.

The rest of categories **Text, Year, Region and Syllables**, were annotated following the same rules as for quantifiers (see section above).[1]

III. MANUAL ANNOTATION

Once the query was run the program CorpusSearch created a file with the extension .cod. This file was used to manually correct the initial searches.[2] There were two categories that needed to be manually corrected: agreement and syllables.

Agreement: both strong adjectives and quantifiers had the same inflectional endings in Old English. The following table (based on [1] and [2]), shows the expected endings of the strong adjective/quantifier if they had been inflected:

	Old English			Early Middle English	Late Middle English
	masculine	feminine	neuter	all genders	all genders
nom.pl	-e	-u	-e/-a	-e	
acc.pl	-e	-u	-e/-a	-e	-e
gen.pl	-ra	-ra	-ra	-er	
dat.pl	-um	-um	-um	-e(n)	

Each adjective/quantifier was, accordingly, categorised as agreeing or non-agreeing, depending on whether they had the expected plural agreement marker or not.

Additionally, there were several adjectives that had French inflectional morphology (-s) in the plural, so these were assigned a different name (French-agree).

Finally, there were some adjectives that were uninflected. Especially those adjectives that had their roots ending on a vowel, could not add the expected -e plural marker and hence were uninflected. For checking which adjectives had the -e ending in their root we used the *Bosworth Toller's Anglo-Saxon Dictionary Online*.

Syllables: for annotating the the number of syllables of a given adjective, the adjective was searched in the *Bosworth Toller's Anglo-Saxon Dictionary Online*. The root (without any inflectional morphology) was used to determine whether the adjective was monosyllabic or polysyllabic. For example, adjectives such as *colde* and *gude* were categorized as monosyllabic because their root *cold* and *gud* is monosyllabic.

IV. FILE CONVERSION

In order to convert the .cod file, so that they could be process in R and share in a repository, we used the query in codes.q, which takes all the strings in the file .cod and creates a new file .ooo with the strings only (plus the added IDs from the corpus). The repository contains these last files (with and without the IDs). This last file is the one that was used for the data analysis process. For information on that the reader is referred to the file: `ME_Case_Study_Agreement_Stats_Graphs.rmd`

[1] R. M. Hogg and R. D. Fulk, *A grammar of old English, volume 2: Morphology* (John Wiley & Sons, 2011).
[2] J. Algeo and T. Pyles, *The Origins and Development of the English Language.(4th cd.)* (Fort Worth: Harcourt Brace Jovanovich College Publishers, 1993).

[1] In the original corpus search another parameter was also annotated: whether the adjective was intersective or not. This information was not used in this work but since it was annotated it has been left there in case it could be used in future research.
[2] For copyright issues these files cannot be shared, but the file .ooo contains all the annotated string and their corresponding IDs in the corpus, so that anyone with access to the corpus can replicate the data.