# Verb lemmatization and semantic verb classes in a Middle English corpus

**Michael Percillier**
Universität Mannheim
Anglistische Linguistik/Diachronie
L13, 9, 68131 Mannheim, Germany
percillier@uni-mannheim.de

## Abstract

The paper describes the creation of new resources and associated tools in the framework of the research project *Borrowing of Argument Structure in Contact Situations* (BASICS), which investigates the borrowing of argument structures of verbs from Old French (OF) to Middle English (ME). The first resource is a database of ME form-lemma correspondences, on which a lemmatization process is based. This process also identifies French-based verbs and thus enables a first diachronic analysis of their prevalence in ME. The second item discussed is a newly developed method for querying ME verbs according to their semantic classes. The created resources and methods are crucial in the continuation of the research project, and can be applied to annotate further ME corpora and train other tools for the treatment of ME data.

## 1 Introduction

This paper is part of a research project[1] that investigates grammatical change in the language contact situation between Middle English (ME) and Old French (OF) that set in after the Norman Conquest (1066) and lasted until ca 1500. More specifically, the project focuses on the connection between the lexical borrowing of verbs and the transfer of their argument structures (AS) from the source language OF to the recipient language ME.

One of the objectives of the project is to trace the spread of AS from borrowed verbs, i.e. verbs originally from OF, to native verbs, i.e. verbs already part of the English lexicon prior to the language contact situation. To this end, corpus queries of syntactic structures need to be complemented by searches for specific verbs and semantic verb classes. The focus on specific verbs allows for a detailed comparison of native and borrowed verbs, while the focus on semantic verb classes will make it possible to follow the spread of syntactic structures from borrowed verbs to other verbs sharing similar meanings.

The currently available resources, described in Section 2, are geared towards queries of syntactic structures, but not specific verbs, let alone semantic verb classes. In order to fulfill the needs of the project, the existing resources have to be enhanced in two ways: (1) the extension of existing annotation with lemma information for verbs, and (2) a method for determining semantic classes of ME verbs. The implementation of both enhancements is described in this paper, as well as their possible application on a recent study of the French borrowing *please* (Trips and Stein, accepted).

## 2 Currently available resources

The *Oxford English Dictionary* (Proffitt, editor, 2015), abbreviated as OED, serves as a point of reference for the project, not only because it is an authoritative resource on the English lexicon, but also because it contains a wealth of etymological information. Owing to a cooperation in the project with the OED's principal etymologist Philip Durkin, we were able to obtain a list of 2,026 English verbs borrowed from French between 1066 and 1500 based on an explicit query. The verbs in said list constitute the starting point of the project, as they are the loan words whose AS is thus introduced to English and can thereafter extend to other verbs.

The ways in which these loan verbs were used should be verified empirically in a corpus. For ME, the *Penn-Helsinki-Parsed-Corpus of Middle English* (Kroch and Taylor, 2000), henceforth PPCME2, presents the advantage of being syntac-

[1] *Borrowing of Argument Structure in Contact Situations: The Case of Medieval English under French influence* (BASICS).

tically annotated. The corpus consists of 55 texts, totaling ca 1.2 million words, and is divided into four periods: M1 (1150–1250), M2 (1250–1350), M3 (1350–1420), and M4 (1420–1500).[2] The annotation format used is *Penn-Treebank*, which can be queried using the specialized software tool *CorpusSearch* (Randall, 2010). The format uses sets of parentheses to represent the clause hierarchy, as illustrated for Modern English in the example below.[3]

```
( (IP-MAT (ADVP-TMP (ADV Then))
        (NP-SBJ (D the)
                (N child))
        (VBD became)
        (ADJP (ADJR happier)
                (CONJ and)
                (ADJR happier))
        (E_S .)) )
```

At the lowest level of the tree hierarchy, each form is assigned a part-of-speech (POS) tag. Consequently, the annotation format, in combination with *CorpusSearch*, makes it possible to search for specific grammatical properties, such as past tense verbs using the *VBD* tag, or specific forms such as *became*. However, due to frequent spelling variation in ME data and the existence of irregular verb paradigms, queries for all forms of a verb, such as *become*, are not readily available by searching for verb stems in ME corpora. To remedy this, all lexical verb forms in the PPCME2 are to be lemmatized, a process described in Section 3.

For the definition of semantic verb classes, the model proposed by Levin (1993), which groups lexical verbs on a semantic basis, can be used as a point of reference. The advantage over other semantic resources such as *WordNet* (Princeton University, 2010) lies in the listing of possible syntactic alternations for each verb class. However, the model applies to Present Day English (PDE) and cannot be directly applied to ME for a number of reasons: (1) semantic changes occurred from ME to PDE, so that the classification proposed by Levin (1993) may be inaccurate for certain ME verbs, (2) ME verbs that no longer exist in PDE are not included in Levin's classification, so that a direct application of the model to ME would result in only partial coverage, (3) a number of PDE verbs did not yet exist in ME and are therefore irrelevant in the definition of ME verb classes, and (4) the potential of syntactic alternations cannot be postulated on the basis of intuition for earlier periods.

In addition to the OED, the *Middle English Dictionary* (McSparran et al., 2001), henceforth MED, constitutes a further dictionary resource that is relevant for the lemmatization of a ME corpus and the definition of ME semantic verb classes. The MED uses unique numerical identifiers (henceforth MED-IDs) for each entry that can serve to disambiguate homonyms. Furthermore, entries in the MED and the OED are linked, so that using both resources in tandem makes it possible to distinguish between native and borrowed ME verbs by checking them against the list of verbs borrowed from French provided by the OED.

## 3 Lemmatization of a ME corpus

As previously stated, the lemmatization of a ME corpus, in particular of its verbs, is a crucial step for any study in which queries of specific verbs or semantic verb classes are to be undertaken. Given the absence of lemmatized ME corpora or any gold standard for the lemmatization of ME data, the lemmatization process relies on the semi-manual assignment of graphemic verb forms to their respective lemmas. The process is divided into two major steps: (1) the creation of an inventory of form-lemma correspondences linking forms in the PPCME2 to lemmas in the MED, and (2) the insertion of this lemma information into the corpus.

### 3.1 Assignment of form-lemma correspondences

Verb forms were extracted from the PPCME2, and each verb form was paired with a lemma and the corresponding ID extracted from the MED. This assignment of verb forms to lemmas was undertaken manually by four trained research assistants and the author using a spreadsheet application. They also had the option of specifying multiple lemmas or marking their choices as doubtful. In total, 19,320 graphemic verb forms were assigned to 2,979 lemmas as primary matches, alongside 4,973 lemmas specified as additional possible matches. The resulting form-lemma links were exported to the YAML (Evans, 2009) format, which was chosen so as to allow the data to be easily imported as

[2] Information from `http://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-4/description.html`.

[3] Example adapted from `http://corpussearch.sourceforge.net/format.html`.

a hash/dictionary in any programming language.[4]

### 3.2 Insertion of lemma information into the corpus

Using the inventory of form-lemma correspondences just mentioned, the insertion of lemma information is performed. For every verb marked with a POS tag beginning with *V* in the corpus,[5] the following instructions are carried out:

The main approach is a lexical lookup in the inventory of form-lemma correspondences. Should this not return any results, two fallbacks are used: (1) Spelling variants are generated and queried for corresponding lemmas. The following grapheme substitution rules are used: $i \rightarrow e/y$, $e \rightarrow i$, $y \rightarrow i/g/+g$, $u \rightarrow v/ou$, $v \rightarrow u$, $th \rightarrow +t/+d$, $+t \rightarrow th$, $+d \rightarrow th$, $g \rightarrow +g/y$, $+g \rightarrow g/y$, $ou \rightarrow u$, $ll \rightarrow l$, $nn \rightarrow n$, and $pp \rightarrow p$.[6] Further, forms containing hyphens or tildes are assigned spelling variants without these characters. (2) The form is stemmed and checked against all stemmed forms in the form-lemma inventory. Stemming is achieved by removing the following ME inflectional suffixes: $+d$, $+d+d$, $+t$, $+t+t$, *an*, *ande?*, *dd?*, *den?*, *e*, $e+d$, $e+t$, *ede?*, *enn?*, *e?st*, *et*, *in?d?e?*, *ingg?e?*, *ode*, *odest*, *oden*, *ten?*, *th*, *tt?*, *yde?*, *ynde?*, *ynn?*, *yngg?e?*, and *yst*.[7]

The lemma information is appended directly to the form in the corpus, so as to still comply with the Penn-Treebank format and related software such as *CorpusSearch*. Each piece of inserted information is demarcated by @ characters and specified by an attribute. Verb lemmas are specified by the attribute *l*, and MED-IDs by the attribute *m* (see Example (1)). For verbs occurring in the list of French-based verbs, an additional attribute *e* (for *etymology*) is defined as *french* (see Example (2)). The attribute *w* (for *warning*) indicates that the lemma was matched using either the spelling substitution or the stemming method (see Examples (2)/(5) and (3) respectively), or that the manual form-lemma match was deemed doubtful (see Example (4)). For verbs spelt as multiple words, the information is appended to the final element (see Example (5)). Should no form-lemma correspondence have been found even after the stemming method, the lemma and MED-ID are marked as *NA* (see Example (6)). The lemma insertion process is summarized in Figure 1.

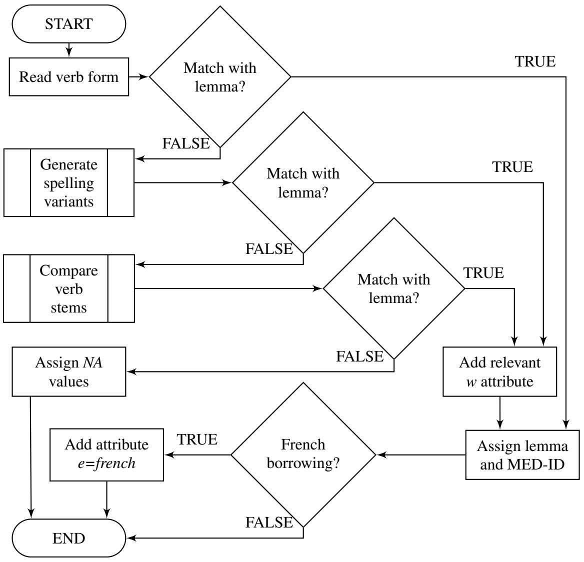

Figure 1: Lemma insertion process.

(1) `(VAG settyng@l=setten@m=39654@)`

(2) `(VAG consyderyng@l=consideren@m=9387@e=french@w=substitution@)`

(3) `(VB tellyn@l=tellen@m=44693@w=stemming@)`

(4) `(VBI wilne@l=wilnen@m=52815@w=doubt@)`

(5) `(VBP21 vnder)(VBP22 stont@l=understonden@m=48362@w=substitution@)`

(6) `(VAN iii@l=NA@m=NA@)`

With this additional annotation, the PPCME2 can be queried for syntactic structures as before, but also for specific verbs. Using *CorpusSearch*, this is achieved by specifying the lemma with the *exists* function, e.g. `(*l=setten@* exists)`. To distinguish between homonyms, the MED-ID can also be used for unambiguous queries, e.g. `(*m=39654@* exists)`.

### 3.3 Evaluation

The lemmatization of verbs in the PPCME2 treated 130,282 verbs in total. 110,116 verbs (84.52%) were directly assigned matching lemmas. Additionally, 5,868 verbs (4.5%) were assigned a lemma

[4] For example with the *PyYAML* (Simonov, 2014) module in *Python* (Python Software Foundation, 2015).

[5] Lexical *be*, *do*, and *have* need not be lemmatized as their tags (*B**, *D**, and *H** respectively) already reveal their lemma.

[6] In the PCCME2, the character sequences $+d$, $+g$, and $+t$ represent the graphemes <ð, ʒ, þ> respectively.

[7] Question marks refer to the regular expression quantifier specifying that the preceding character may or may not occur.

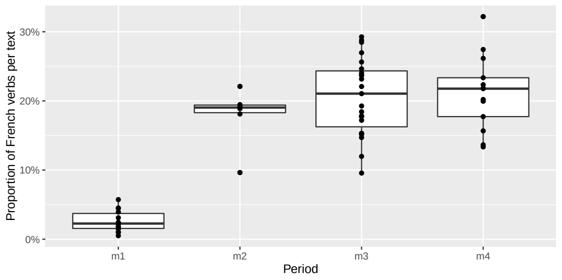

Figure 2: Proportion of French-based verbs in sub-periods of ME.

| | M1 | M2 | M3 |
|---|---|---|---|
| **M2** | 1.0e-07 | - | - |
| **M3** | 8.2e-15 | 1 | - |
| **M4** | 2.6e-13 | 0.97 | 1 |

Table 1: Pairwise t-tests of French-based verbs per ME sub-period, using Bonferroni correction.

using spelling substitution, and 10,421 verbs (8%) using stem comparison. The total of lemmatized verbs is thus 126,405 (97.02%), whereas 3,877 verbs (2.98%) could not be assigned any lemma. Based on controls of random samples of 100 tokens, the spelling substitution and stem comparison fallbacks were estimated to be accurate to 86% and 90% respectively.

The estimation of French-based verbs and the division of ME into the sub-periods M1–M4 make it possible to investigate the diachronic spread of French-based verbs in ME (see Figure 2).[8] The analysis suggests a strong increase in the usage of French-based verbs between M1 and M2, with only little fluctuation thereafter. This is confirmed through pairwise t-tests (see Table 1) with Bonferroni correction (Baayen, 2008, 105–106).

## 4 Determining ME semantic verb classes

In order to identify ME semantic verb classes, the classification proposed by Levin (1993) can serve as a point of reference, but cannot be applied directly to ME, as already mentioned in Section 2. The estimation of ME equivalents to the semantic verb classes proposed by Levin (1993) is undertaken in three steps: (1) the creation of a database of semantic classes and the verbs therein from which verb lists can be extracted, (2) a method for finding ME verbs synonymous to the PDE verbs extracted in the previous step, and (3) a method for querying the corpus for multiple verbs simultaneously.

### 4.1 Creating an inventory of semantic verb classes

An electronic index of Levin (1993) exists as a HTML file,[9] but it only lists which verbs occur in which numbered section of the monograph, thereby omitting the names and descriptions of the classes entirely. An updated index was therefore generated that not only numbers but also names classes. This index can be queried with a script that parses the HTML tree[10] and allows for two types of searches: (1) by verb class, to determine which verbs occur in a given class, or (2) by verb, to determine to which verb classes a particular verb belongs.

### 4.2 Matching ME and PDE verb meaning

Determining ME equivalents to PDE verb classes proposed by Levin (1993) entails finding ME verbs synonymous to verbs listed in PDE classes. The MED allows a “reverse lookup” of ME verbs via its search engine[11] when specifying a PDE verb as a query within entry definitions, which returns a list of MED entries in which the query term occurs anywhere within the definition. For example, a reverse search for the PDE verb *acknowledge* returns ME verbs such as *agraunten* (‘to acknowledge, grant’), *aknouen* (‘to recognize (sth.) as a fact, acknowledge, know’), or *kithen* (‘to acknowledge (sb.) as (sth.)’). A script automates the process by querying a given list of PDE verbs, then excluding any results that are not verbs.

This list of verbs requires manual verification for two reasons: (1) the presence of a verb in the list merely indicates that the query item was found within the definition, but not necessarily that the PDE and ME verbs are synonyms, and (2) the PDE verb used in a query may be polysemous, so that the PDE and ME lemmas may be correctly matched, but their specific meanings may differ. An example of the first point is a query of the PDE verb *crown* that returned the ME verb *cacchen* (‘to catch’) because the MED entry contains “**cacchen of**: take off (one’s crown, etc.) quickly”. In this case, the matched string referred to the noun *crown* as used in an example sentence, and therefore does not

[8] Figure generated in R (R Core Team, 2016) with *ggplot2* (Wickham, 2009) and *scales* (Wickham, 2016).

[9] `http://www-personal.umich.edu/~jlawler/levin.html`

[10] Using the module *BeautifulSoup* (Richardson, 2015).

[11] `http://quod.lib.umich.edu/m/med/structure.html`

constitute a valid match. The second point can be illustrated by the PDE verb *consider*, which is listed as a "verb with predicative-complements" by Levin (1993, 181), more specifically in the subclass "*appoint* verbs". However, this classification only applies to a specific meaning of *consider*, i.e. "to regard in a certain light or aspect", but not to other meanings such as "to think/contemplate". The separation of different meanings of polysemic verbs is crucial, given that they result in distinct AS (Löbner, 2002, 114–116). For this reason, valid matches for PDE verbs have to be checked for congruence with the specific meaning used in a given semantic class. ME verbs that fulfill these conditions can be considered as semantic equivalents to the PDE class defined as input for the query.

### 4.3 Querying multiple verbs

Simultaneous queries of multiple verbs can be specified in a `*.q` file that serves as input to *CorpusSearch*. The query language allows the logical operator `|` (OR), so that multiple MED-IDs can be searched, e.g. `(*m=9348@*|*m=9356@* exists)`. As the lists of ME verbs to be queried can be long, the creation of such query files is automated via a script that reads the list of MED-IDs from a column named "MED-ID" in a CSV table, then generates a corresponding `*.q` query file.

### 4.4 Application of the method

The proposed method of identifying ME semantic verb classes and querying the verbs in a simultaneous manner has direct applications for recent and ongoing studies.

For instance, Trips and Stein (accepted) empirically verified the assumptions proposed by Allen (1995) on the transfer of prepositional 'datives' from the French-based verb *plesen* ('to please') to the native verbs *liken* ('to like') and *quemen* ('to please'). They conclude that ME, having lost most of its formal case distinctions, adopted the 'dative' arguments of the donor language OF. The semantic properties of the borrowed verb *plesen* allowed the transfer of its AS, specifically the use of prepositional objects, to native verbs belonging to the same semantic class of verbs of psychological state, so-called *psych* verbs (Levin, 1993, 188–193). This transfer led to a rise in the use of prepositional objects with native psych verbs, with *quemen* ultimately replaced by *plesen*. The new structure eventually spread to native verbs belonging to other semantic classes, e.g. *yeven* ('to give').

The important findings presented by Trips and Stein (accepted) regarding the ME verbs *liken*, *quemen*, and *plesen* can be systematically verified for other ME psych verbs by using the proposed method of identifying ME semantic verb classes. Furthermore, the spread of new structures to verbs of other semantic classes, as in the case of *yeven*, can also be investigated by examining whether certain semantic classes adopted the new structures more frequently or more quickly than others.

## 5 Conclusions

The present paper discussed two enhancements to a parsed corpus of ME that are necessary for a project investigating AS borrowing from OF to ME. The first enhancement is the lemmatization of lexical verbs, so that queries for specific verbs can be performed in addition to searches for syntactic structures. The second enhancement builds upon the first in that it allows for the search for multiple verb lemmas at once, more specifically those belonging to a given semantic class. By identifying French-based verbs, the lemmatization process also enabled a diachronic analysis of the proportion of French-based verbs per ME sub-period.

## 6 Outlook

The two processes discussed are vital for the project at hand, as they clear a methodological "bottleneck", thus allowing searches for specific ME verbs and semantic classes to proceed. Furthermore, the analysis of the proportion of French-based verbs raises an interesting research question pertaining to the delay between the wider adoption of French-based verbs in M2 and the expansion of their AS to native verbs.

Although tailored to a specific project, the resources and methods have further applications. The form-lemma links and their lemmatizer script can be applied to other ME corpora, and a general lemmatizer for ME (i.e. not limited to verbs) can benefit from this inventory as a training resource. The index of PDE semantic verb classes and the method to adapt it to ME can be used to perform semantic verb class searches in both ME and PDE corpora.

## 7 Resources

The created resources and their associated tools are available via the *BASICS Toolkit* web application.[12]

[12] `http://terrano.philosophie.uni-stuttgart.de/BASICStoolkit`

## Acknowledgements

Funding from the *Deutsche Forschungsgemeinschaft* (grant TR555/6-1) is gratefully acknowledged. I thank Carola Trips, Achim Stein, Yela Schauwecker, and Richard Ingham for their cooperation in the BASICS project. For their work on the form-lemma correspondences, I thank Lena Kaltenbach, Natascha Schultheiß, Lisa Seidel, and Jonas Stork. I would also like to thank the three anonymous reviewers for their useful suggestions.

## References

Cynthia L. Allen. 1995. *Case Marking and Reanalysis: Grammatical Relations from Old to Early Modern English*. Oxford University Press, Oxford.

Rolf Harald Baayen. 2008. *Analyzing Linguistic Data: A Pratical Introduction to Statistics Using R*. Cambridge University Press, Cambridge.

Clark C. Evans. 2009. YAML. `http://yaml.org`.

Anthony Kroch and Ann Taylor. 2000. The Penn-Helsinki Parsed Corpus of Middle English, Second Edition (PPCME2), Release 3. `http://www.ling.upenn.edu/hist-corpora/`.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago.

Sebastian Löbner. 2002. *Understanding Semantics*. Routledge, London.

Frances McSparran, Paul Schaffner, John Latta, Alan Pagliere, Christina Powell, and Matt Stoeffler. 2001. Middle English Dictionary. `http://quod.lib.umich.edu/m/med/`.

Princeton University. 2010. WordNet. `http://wordnet.princeton.edu`.

Michael Proffitt, editor. 2015. Oxford English Dictionary. `http://www.oed.com`.

Python Software Foundation. 2015. Python 2.7.10. `https://www.python.org`.

R Core Team. 2016. R: A language and environment for statistical computing. `https://www.R-project.org/`.

Beth Randall. 2010. Corpussearch 2.003.00. `http://corpussearch.sourceforge.net`.

Leonard Richardson. 2015. Beautiful Soup Version 4.41. `http://www.crummy.com/software/BeautifulSoup/`.

Kirill Simonov. 2014. PyYAML. `http://pyyaml.org/wiki/PyYAML`.

Carola Trips and Achim Stein. accepted. Contact-induced changes in the argument structure of Middle English verbs on the model of Old French. *Journal of Language Contact*.

Hadley Wickham. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer, New York.

Hadley Wickham. 2016. scales: Scale functions for visualization. `https://CRAN.R-project.org/package=scales`. R package version 0.4.0.