Univerzita Karlova v Praze

Filozofická fakulta

Ústav anglického jazyka a didaktiky

Filologie – Anglický jazyk

Ondřej Tichý

Nástroj na tvaroslovnou analýzu staré angličtiny

Morphological Analyser of Old English

Teze k disertační práci

vedoucí práce - Prof. PhDr. Jan Čermák, CSc.

2014

Contents

Introduction

The thesis describes a project aiming to produce the Morphological Analyser of Old English, a computer program that receives Old English (OE) text on input, automatically analyses it and outputs all morphological functions that the forms in the text can conceivably carry[1].

[1] Based on their morphology and within our understanding of OE grammar.

Aims

The motivation to create the program was manifold, but mainly pedagogical and scientific in nature.

The program should allow students to query comprehensive lexicographical resources without prerequisite knowledge of OE morphology, dialectology or medieval spelling conventions.

Teachers, on the other hand, often need to compile extensive glossaries for OE texts and a tool that could generate a basic glossary requiring only some fine-tuning by the teachers would save them precious time.

But, most importantly, diachronic linguistics has lately been more and more driven by corpus methodology. With inflectional languages, much of this methodology relies on the existence of lemmatised corpora. Since there is no lemmatised corpus of OE,[2] the corpus research of OE has been severely limited. The tools for lemmatisation of OE in existence today are more or less unusable and the dissertation attempts to make the first step to change this – by introducing a tool for semi-automating the lemmatisation process.

Structure

The thesis examines the state of the art in machine morphological analysis of OE as well as the successful implementations of such tools in a structurally comparable language (Czech), while noting features applicable to OE on the one hand and the difficulties previously not encountered and specific for OE on the other hand. It describes the OE inflectional morphology with regard to the requirements of this project and it characterizes OE spelling and variation as the major obstacles of the project. After an overview of the technology used by the project, the paper then describes the implementation of the program itself in three major parts: 1. the processing of the input data; 2. the generation of inflected forms; and 3. the text analyser proper. The dissertation concludes with a discussion of the results (of the automatically analysed texts) and suggestions for further improvements and development.

[2] Of any practical size – see chapter 1.3. of the dissertation for a description of several small sized manually lemmatised corpora.

Theoretical Framework

The survey of the current state of the art in the field of programmatic morphological analysis of OE concludes that the methods used so far are not suitable for the present project. The methods are either inadequate in view of the aims or they would require resources beyond our capacity – and often both.

Therefore, the plan of implementation is not based on existing work in the field of OE, but on automatic morphological analysis of Modern Czech as implemented by Osolsobě (1996), Sedláček (1999) and Sedláček & Smrž (2011).

The framework consists of three major phases:

1. Preparation of the **input data** (the lexical data and the information about the lexical items necessary for phase 2).
2. Creation of a dictionary of inflectional forms from the data prepared in phase 1 by an automatic **generator**.
3. Matching items from an OE text with items in the dictionary of forms by the **analyser** proper.

The sources of the input data and the specific operation of each phase is based on an overview of OE inflectional morphology and of its variation.

Old English Inflectional Morphology

The thesis does not cover the whole OE inflectional system in detail, but only describes those parts of the structure necessary for the operation of the analyser.

In doing so, it diverges from the traditional categories and paradigms that originate in historical linguistics and whose aim is to show the continuity of OE as an offspring of the Germanic languages. At the same time, it also diverges from more recent simplified descriptions aimed at contemporary students.

Instead, the level of detail of description has been chosen pragmatically – e.g. a lower one for adjectives, a higher one for verbs. The unifying principle was efficiency in view of the computational processing and feasibility.

For example, if the differences between some traditional inflectional paradigms are smaller than the variation or oscillation[3] of their members between the paradigms, or if a merger of the paradigms does not cause problems in analysing OE texts, the paradigms are merged in this description. If, on the other hand, the number and character of exceptions from a paradigm warrants creation of new paradigms, those paradigms are established. It is important to remember that due to the requirements of the project, the description is "dictionary-form oriented" (see Implementation), i.e. the paradigms are defined so that with the dictionary forms, with a set of endings and with as few replacement rules as possible it is easy to inflect any member of a paradigm.

In the case of nouns and adjectives (and classes inflected accordingly), one paradigm is usually defined per set of endings. In the case of verbs, a paradigm is defined for every set of infixes and endings.

However, if there are no special reasons against it, Wright's and Campbell's descriptions are closely adhered to, because they provide extensive lists of words assigned to their paradigms and because it may be useful to be able to refer users of the analyser back to the traditional and widely accepted reference books for further details on forms and functions. Major departures from these traditional descriptions are therefore noted in the thesis.

Any project attempting to process an authentic non-standardized OE text computationally has to deal with the problem of variation (scribal, dialectal or diachronic in nature) and it is important to understand the nature of the variation of the vernacular on the one hand and the ways modern reference books choose to standardize it on the other hand.[4]

[3] Be it scribal variation, variation based on sound change or analogical shifts in class paradigm membership.

[4] Especially if we base our project on the modern reference books, as explained below, and not only on the vernacular of the day.

The thesis explores the major types and sources of variation and recognizes those types of variation that can be safely ignored and standardized on input – i.e. during the first phase mentioned above (e.g. certain types of allography); or safely ignored but preserved for the output for the benefit of the users (e.g. vowel length); or those that cannot be ignored and have to be dealt with either in the second or the third phase of the analysis (e.g. length of consonants or any other type of contrastive grapheme variation).

Implementation

Both technology and the implementation had been determined (beside the aims defined in the Introduction) by the means at our disposal (especially technological and personal).

The personal limitations preclude a large-scale preparatory stage requiring a team of OE experts that would either tag a training corpus or manually inflect all the attested OE lexical items (i.e. create a complete dictionary of inflected forms). The only other option available (barring some kind of a purely statistical approach) involves using an existing lexical source and a linked source of grammatical information to produce inflected forms that match with the text submitted by the user.

The technological limitations dictate that we pre-generate forms so that the matching process itself is as simple and therefore as fast as possible even on low-grade hardware at our disposal.

Technology

Three factors contributed to the choice of the technology:

1. The project is built on free and open-source technology. Since all the input data are to be public-domain, all results of the project can (and will be) made publicly and freely available.
2. The technology employed for the project should be widely used allowing for future modification of the project by others, as well as for multiplatform deployment so that it can reach as wide an audience as possible.

3. The tools should be adequate for the project goals they are to accomplish. That means they should be easy to use in achieving these goals both by the authors and by the users. This especially entails their efficiency in dealing with the functions and data in question.

4. The generator script is programmed in PERL 5.16. PERL is a high-level programming language well-known for its qualities in text manipulation and commonly used by linguists. It is a natural choice for a script whose major functions are loading, transforming and storing strings of texts. PERL's strong implementation of regular expressions used here to match and replace strings is one of the important factors for its choice as is its support of the Unicode standard.

5. To promote the online use of the project, the exported data are also stored in a MySQL database. MySQL is a relational database allowing data to be queried more efficiently than in the case of a simple text file. The tables of the database run on both the MyISAM and the InnoDB engines.

6. The analyser script is programmed in PHP – a server side scripting language widely used for dynamic web content and for processing database queries, often in concert with the MySQL databases. It is in many ways similar to (and partly derived from) PERL and can thus make use of some existing procedures already used in the generator. Also, it should be relatively easy to transform the script into PERL if an offline analyser should ever be needed (e.g. for annotating large corpora).

7. The user interface is in HTML with some JavaScript used for client-side scripting (e.g. for inserting special symbols by mouse or displaying reference sources on mouse-click).

Input

The thesis describes the implementation step by step in a succession determined by the phases mentioned above.

Based on a survey of existing sources of lexical and grammatical information on OE, the online version of *An Anglo-Saxon Dictionary* by Bosworth and Toller (*BT*) was chosen as the most comprehensive as well as the most accessible source of lexical information. For grammatical information Wright's *Old English Grammar* was chosen as the best matching source.

A large amount of preparatory work had already been carried out on the *BT* before the start of the current project (Tichý, 2007). Additional work on both the macrostructure and the microstructure of the *Dictionary* was necessary before the data could be exported for the purposes of the analyser.

Wright's *OE Grammar* has been used to improve the quality of the grammatical information in the online dictionary. For this reason, it is partially present in the data exported from the *Dictionary*. Namely, each headword of the *Dictionary* identifiable in Wright's index has been associated with the corresponding paragraphs of the grammar book.

Due to the more complicated system of verbal conjugation, a dictionary of verbal paradigms has been manually constructed. This dictionary allows for the generation of verbal forms including infixation in the next phase based solely on base forms of lexical items.

Generator

The lexical data and the dictionary of verbal paradigms are loaded and processed by the generator (standardised to a degree and enriched by automatically calculated information necessary for the following operation of the generator, e.g. syllable count or stem weight).

The generator then operates in two main stages. First it assigns each lexical item a paradigm and then it generates all inflectional forms based on the base form and the associated paradigm.

Paradigm Assignment

The assignment of paradigm examples (or the paradigms they represent) to individual lexical items is carried out in several stages, separately for each word-class. The order of the stages is given mainly by the reliability of information upon which the decisions about the paradigm affiliation can be made.

The stages of the paradigm assignment are as follows:

1. With verbs, the "stems" of the lexical items are string-compared to the lemmata of the paradigm examples and the verb type of the lexical items is string-compared to the verb type of the paradigm examples.[5] If both match, the paradigm is assigned.
2. The first step is improved on by going through all the unassigned verbs and string-comparing the beginnings of their "stems" with items on the list of verbal prefixes (derived from *BT*). If there is a match and the rest of the "stem" is then successfully compared to a paradigm example lemma, the verb is assigned. This step is useful, since *BT* is not quite consistent in marking prefixes and suffixes with hyphens, especially if there is more than one prefix in existence.
3. When stem comparison can yield no more results, the unassigned lexical items with links to Wright's paragraphs are assigned paradigms corresponding to those paragraphs.

   Since the other inflected word-classes beside verbs do not have an external set of paradigms, this is the first step in which their lexical items are assigned. The comparison is therefore not with paradigm examples loaded from a file, but

[5] This prevents some homonyms to be mistakenly assigned.

each lexical item is searched for one (or more) paragraph number(s) and assigned a corresponding paradigm directly in the source code.

4. With a group of newly assigned items, the stem comparison from step 2 is repeated. In this way, even items derived from those mentioned in Wright, but not present in his index themselves, can be assigned.
5. At this point, the more reliable information based on Wright and stem similarity is exhausted and the stems themselves are analysed for phonological or morphological clues as to their paradigm affiliation. This obviously differs for each word class. Due to the nature of the algorithm, all adjectives should be now assigned.
6. With each group of the newly assigned items, the stem comparison from step 2 and 4 is repeated, but this time a certain degree of expected variation is allowed.
7. The rest of the unassigned items is assigned according to the following rules:
    a. All the unassigned strong verbs are assigned to the *helpan*-type; the rest of the verbs is then assigned to the *déman*-type.
    b. All the unassigned masculine nouns and all the unassigned nouns of uncertain gender are assigned to the *stán*-type, all the feminine nouns are assigned to the *ár*-type and all the neuter nouns are assigned according to their stem length. Neuter long stems follow the *stán*-type, short stems follow the *hof*-type.

Form Generation

Once the lexical items are assigned their respective paradigms, forms are generated, word-class by word-class, mostly by concatenating stems and endings. In the case of verbs, the stems are concatenated themselves according to the dictionary of verbal paradigms. The endings and their variants are supplied directly in the generator script.

Each generated form is supplied with the respective grammatical information as well as with some information on its composition (e.g. its structure, paradigm, class etc.).

The endings are based on Wright and on the dialectal data from Campbell. This way, some of the more predictable and prevalent variants are pre-generated together with standard forms and make the following matching of the forms simpler.

Since some inflectional forms (like personal pronouns) are impossible (or rather not effective) to generate, these were defined manually and are simply added to the forms generated automatically.

All the forms thus obtained are then outputted by the generator and loaded into an online database to be used by the analyser script in the following phase.

Analyser

The aim of the online analyser script is to:

1. take OE text from the end-user on input,
2. process it to derive a list of all individual types,[6]
3. match each type with a form in the database,
4. fetch all data about the form from the database,
5. and, finally, present the data about each form back to the end-user.

The crux of the whole operation of this project lies in step 3 and is complicated by the aforementioned variation. Some of the variation poses no problem, since variant forms were generated in the previous phase, but many instances of variation are either difficult to predict, or ineffective to be generated.

For that reason, forms are not matched by a simple string comparison only, but also through the “variation filters” that modify the matching strings if a direct match is not accomplished. The variations are loosely based on Baker (Common Spelling Variants, 2012) and Wright, but some were combined into one or simplified. The string modification using the variation filter proceeds from simple substitution (e.g. *u* for *v*) to

[6] By type we here mean a unique word-form, not a lemma.

complex regular expressions that may allow e.g. a variation of any back vowels in front of a nasal.

All matched forms are displayed under the particular type from the input text. Currently, the forms are grouped by word-classes and lemmata in a glossary-style output.

Results

As expected, a certain portion of the input text cannot be currently matched (i.e. analysed) by the program. The dissertation discusses results obtained from processing 10 OE texts of various provenance (comprising ca. 2 500 words) by the Analyser and concludes by establishing the current recall rate at 95%[7], with suggestions on how to improve the rate especially for dialects other than West Saxon.

Conclusion

The thesis concludes with a discussion of the future use and development of the analyser.

The major goal beside partial improvements on all the levels of the current program is the development of a disambiguator, since at this stage, the analyser returns all morphologically feasible results, as was originally planned. Due to the variation and especially grammatical homography/homonymy of the already weakened OE inflectional morphology, the ratio of lemmata returned by the analysis for a particular type is very high – on average 3:1 –, and the ratio of form-function pairs for a particular type is accordingly higher – on average almost 10:1.[8]

If one of the main goals (i.e. lemmatisation of OE corpora) of the long term project this thesis forms only a part of is to be achieved, the results of the analyser will have to be disambiguated. The thesis makes some suggestions on how to improve the analyser in

[7] The recall is as high as 100% for texts most similar to the standard described in modern reference books (i.e. for texts in West Saxon dialect of the ca. 10th century), but it is almost 5-10% lower for Northumbrian or even 10-15% lower for Kentish texts.

[8] The high number of lemmata is mostly due to the problems in the macrostructure of the *BT* and some grammatical homonymy of e.g. imperative sg. forms of derived verbs and the base forms of the corresponding nouns or adjectives. The high number of form-function pairs is mostly due to high grammatical homography, or, in other words, due to a small number of morphological forms available for a high number of functions (e.g. in adjectival inflection).

order to give preference to some of the results of the analyser in the following disambiguation, but it is clear that the process will have to be based on other than morphological principles as well, if not to a major degree.

The aims (as specified in the Introduction) were achieved, though the precision of the analyser needs to be further improved and the technical requirements for making the tool public (e.g. as part of the online *BT*) need yet to be resolved.[9] Only a continued use by students, teachers and scholars will show, how effective the tool is in each of its applications.

[9] These are mainly security and performance requirements.

Bibliography

Adams, J. (2007). Retrieved from A Morphological Analyzer for Old English Verbs: http://www.cs.cmu.edu/~jmadams/NLPLabProject.pdf

Baker, P. S. (2012). Common Spelling Variants. Retrieved from The Electronic Introduction to Old English: http://www.wmich.edu/medieval/resources/IOE/variants.html

Baker, P. S. (2012). Pronunciation. Retrieved from The Electronic Introduction to Old English: http://www.wmich.edu/medieval/resources/IOE/pronunciation.html

Baker, P. S. (2012). Verbs. Retrieved from The Electronic Introduction to Old English: http://www.wmich.edu/medieval/resources/IOE/inflverb.html

Bosworth, J. (1921). An Anglo-Saxon Dictionary: Based on the Manuscript Collections of the Late Joseph Bosworth. (T. N. Toller, Ed.) Oxford: Clarendon Press.

Calle Martín, J., & Triviño Rodríguez, J. L. (1998). Algoritmos de derivación de palabras con ortografía irregular en el análisis morfológico automático del Inglés Antiguo. Interlingüística, 67-70.

Campbell, A. (1983). Old English Grammar. Oxford: Clarendon Press.

Čermák, J., & Znojemská, H. (2001). Čítanka staroanglických, středoanglických a raně novoanglických textů. Praha: Nakladatelství Karolinum.

Hajič, J. (2004). Disambiguation of Rich Inflection. Prague: The Karolinum Press.

Hall, J. R. (1894). A Concise Anglo-Saxon Dictionary. London: Swan Sonnenschein & Co.

Healey, A. d. (Ed.). (2014). Dictionary of Old English A to G Online. Retrieved from http://www.doe.utoronto.ca/

Healey, A. d., Holland, J., McDougall, I., & Mielke, P. (2000). The Dictionary of Old English Corpus in Electronic Form. Toronto: University of Toronto.

Hogg, R. M., & Fulk, R. D. (2011). A Grammar of Old English: Morphology. Wiley-Blackwell.

Kay, C., Edmonds, F., Roberts, J., & Wotherspoon, I. (2005). A Thesaurus of Old English. Amsterdam: Rodopi.

Kiernan, K. (2011). Electronic Beowulf. London: British Library.

Kleinman, S. (20012). A demo version of the OE lemmatiser. Retrieved from http://www.csun.edu/english/lemmatise/main.php

Lindberg, T. (2014, February 7). Retrieved from Verbix: http://www.verbix.com/

Martín Arista, J. (2014, February 7). Retrieved from Nerthus Project: http://www.nerthusproject.com/

McGillivray, M. (2014). Cynewulf and Cyneheard. Retrieved from Old English Texts: http://www.ucalgary.ca/UofC/eduweb/engl401/texts/ohthfram.htm

McGillivray, M., & Chevallier, G. (2014). The Voyage of Ohthere. Retrieved from Old English Texts: http://www.ucalgary.ca/UofC/eduweb/engl401/texts/ohthfram.htm

Miranda García, A., Calle Martín, J., Moreno Olalla, D., & Muñoz González, G. (2006). The Old English Apollonius of Tyre in the light of the Old English Concordancer. In A. Renouf, & A. Kehoe, The Changing Face of Corpus Linguistics (pp. 91-98). Rodopi.

Miranda García, A., Triviño Rodríguez, J. L., & Calle Martín, J. (2000). MAOET: Morphological Analyser of Old English Texts. Proceedings of the 10th International Conference of SELIM (pp. 127-145). Zaragoza: Institución Fernando el Católico.

Mitchell, B., & Robinson, C. F. (2001). A Guide to Old English (6th ed.). Oxford: Blackwell Publishing.

Osolsobě, K. (1996). Algoritmický popis české formální morfologie a strojový slovník češtiny (unpublished disertation). Brno: Masarykova Univerzita.

Oxford University Press. (2014). OED Online. Retrieved from http://www.oed.com/

Quirk, R., & Wrenn, C. L. (1957). An Old English Grammar. London: Routledge.

Rissanen, M., Kytö, M., Kahlas-Tarkka, L., Kilpiö, M., Nevanlinna, S., Taavitsainen, I., . . . Raumolin-Brunberg, H. (1991). The Helsinki Corpus of English Texts. Helsinki: University of Helsinki.

Sedláček, R. (1999). Morfologický analyzátor češtiny (unpublished diploma thesis). Brno: Masarykova Univerzita.

Sedláček, R., & Smrž, P. (2001, June). Automatic Processing of Czech Inflectional and Derivative Morphology. FI MU Report Series.

Sweet, H. (1887). A Second Anglo-Saxon Reader: Archaic and Dialectal. Oxford: Calendon Press.

Taylor, A., Warner, A., Pintzuk, S., & Beths, F. (2003). The York-Toronto-Helsinki Parsed Corpus of Old English Prose.

Taylor, A., Warner, A., Pintzuk, S., & Plug, L. (2001). The York-Helsinki Parsed Corpus of Old English Poetry.

Tichý, O. (2007). Digitization of Old and Middle English Dictionaries (unpublished thesis). Prague: Faculty of Arts, Charles University in Prague.

Wright, J., & Wright, E. M. (1914). Old English Grammar . London: H. Milford, Oxford University Press.