An Anglo-Saxon Dictionary and a Morphological Analyzer of Old English

Ondrej Tichy
Jan Čermák
Charles University in Prague

The main stages in the project of the digitization of the Anglo-Saxon Dictionary *by J. Bosworth and T. N. Toller are described and the value of the resulting data is considered. The paper suggests that the dictionary data need to be structurally tagged if we are to further benefit from the project beyond the current dictionary application. It is also noted that the re-tagging process can be partially automatized, but that it will have its complications due to the ambiguity of typographical tagging currently included in the data. An outline of the development of an Old English morphological analyzer, now in its early stages, is offered using the valuable digitized data of the Dictionary and drawing on a model of a functional Czech morphological analyzer. Envisaged problems, such as the building of stem- and affix-lexicons, Old English vowel variation and stem-final variation, are discussed and several solutions are proposed. The paper also proposes and accounts for some divergence from the model of the Czech analyzer reflecting differences between Czech and Old English morphology and slight differences in the final uses of the Modern Czech and Old English analyzers. Finally, the analyzer's future use, both as a part of the dictionary and as a stand-alone tool for parsing the corpora, for connecting the lexicon entries with text, etc., is suggested and some possibilities of future improvements, e.g. a word-formation or a syntactic analyzer, are indicated.*

Introduction

An Anglo-Saxon Dictionary by Joseph Bosworth and T. Northcote Toller (BT) has been the most complete dictionary of Old English for over a hundred years now and, with the exception of the long-run project of the DOE,[1] its primacy has not been challenged yet. As such it has been a primary resource for Anglo-Saxonists—historians, medievalists and historical linguists alike—since its inception. It may therefore seem rather surprising that such a unique tool is practically unavailable or at least difficult to reach for many of its potential users.[2]

However, the original project of the digitization, as conceived and directed by Sean Crist under his Germanic Lexicon Project (GLP), was not primarily motivated to serve medievalists in need of a translator's dictionary. The GLP in general aimed "to create comprehensive electronic documentation of the lexicons of the early Germanic languages, particularly of the etymological relationships among the words in those languages" (Crist 2001), but it has been obvious from the beginning that if the digitization project succeeds "[d]ifferent scholars will be able to use this resource for different purposes" (ibid.).

It was with a distinct goal to create an easy to use and easy to get version of the BT dictionary—not least for medievalists in need of such a tool—that we joined the GLP in 2005. Only in the process of accumulating the data did we realize that the dictionary data might have

[1] "The Dictionary of Old English developed by the Centre for Medieval Studies at the University of Toronto is based on a complete corpus of Old English aiming to be an exhaustive dictionary of OE, thus surpassing BT. It was initiated in 1969 and currently has about half of its entries finished." (DOE, 2006)

[2] The book has been out of print for a long time and though its text is in public domain, its market price is well over £300.

different exciting applications besides those of its printed original and that these applications might require a variety of solutions we did not envisage in the beginning.

This paper will describe 1. the basic digitization process that we have participated in under the GLP; 2. our subsequent and independent processing of the acquired data; and 3. our plan for the morphological analyzer, which has ensued from the previous.

Digitization

The digitization was initiated by Sean Crist in 2001, when the dictionary was scanned and a basic text was generated by OCR software from the images. At this stage, a preliminary general analysis was carried out so that the most frequently recurring errors could be automatically corrected.

After this, the most time-consuming part of the project followed: more than 2000 dense, large-format pages of the dictionary had to be hand-corrected. This started as a voluntary enterprise, but, thanks to the John Hus Educational Foundation grant, it has been possible for our team to join the project actively and to accelerate the process.

The decision to store the text in plain ASCII[3] and to encode any non-ASCII characters by an extended set of HTML/XML entities with formatting in standard HTML tags proved a wise one—new, unanticipated characters kept springing up during the correction process well until its end, many of them unknown even to the Unicode standard and possibly never depicted in an electronic font before.

A preliminary version of a dictionary application was created by our team[4] in May 2007 and the digitized text was also incorporated into a simple on-line full-text search engine developed by Sean Crist for the GLP.[5] These should provide users with a free and an easy to use version of the dictionary, but there is not much of added functionality compared to the paper version of the dictionary. Faster browsing or a full-text search (with some wildcard support) can obviously be helpful, but there is much more that an electronic dictionary can do, like discrete search through individual entry elements (equivalents, etymologies, examples, references, etc.), user defined views, automatic lemmatizer of users' input or a complete morphological analyzer.

Tagging

To accommodate any of these functions, the dictionary has to be first and foremost structurally tagged. The paper dictionary makes use of several typographical features to structure its entries, but as the set of these features is much smaller than the number of different micro-structural elements, the features are necessarily ambiguous. Thus italics is used for both the Modern English equivalents, Latin glosses and grammatical information; bold signifies the head of the entry but it can also mark important variants, subentries or references to other main entries. Many features are distinguished solely by position, namely the quotations, their sources and dates. In fact, the typography of the original dictionary is quite ingenious (considering the limited variety of formatting it had at its disposal), though perhaps not as developed as that of the NED. Still, as it is, it has only a limited use for automatic processing.

The re-tagging process can be partially automatized (a set of rules and conditions that should serve as a basis of the re-tagging algorithm has already been formed), but it will have to be

[3] ASCII (American Standard Code for Information Interchange) defines 95 printing characters derived from the English alphabet. Unlike some more advanced encodings, any contemporary PC can be supposed to read and interpret ASCII correctly.

[4] *http://lexicon.ff.cuni.cz/app.*

[5] *http://lexicon.ff.cuni.cz/search/aa_search.html.*

carefully planned and supervised. An "in house" XML style sheet has been devised, but a TEI[6] compliant style sheet is also under consideration.

Morphological analyzer

The most compelling project following from the digitized dictionary promises to be the development of a morphological analyzer. It could greatly improve the user-friendliness of the digitized dictionary (e.g. by lemmatizing the user input or by providing additional grammatical information), but its applications may go far beyond the dictionary project (automatically parsing and tagging Old English corpora, tools for semi-automatic glossary creation, basis for word-formation[7] and syntactic analyzers, etc.).

The principles of the analyzer's operation

As we will try to show now, the essential part of the analyzer has to be based on the lexicographical data of the BT. The model for our analyzer is the automatic morphological analyzer of Czech developed by Sedláček and Smrž (2001) based on an *Algorithmic Description of Czech Formal Morphology and a Czech Machine-Readable Lexicon* (Osolsobě 1996). The comparison of Modern Czech and Old English may seem at first unorthodox, but in our opinion the typological characteristics of Modern Czech and Old English morphology are close enough to justify the use of similar methods in this case.

Two lexicons or "wordlists" will form the heart of the analyzer, the lexicons of: (a) stems and (b) endings (most productive derivational suffixes may be considered as well, but generally only grammatical endings will be included at this point).[8] The input of the analyzer will then try to separate, in each case, a stem and an ending and after processing these with the "filters" described below, it will attempt to identify each of these in the appropriate lexicon, making it thus possible to output the corresponding lemma(ta), grammatical information about the input form, etc.

The lexicon of endings will be created manually, using the standard descriptions of Old English morphology such as Campbell's *Old English Grammar* (1959). Apart from the affixes themselves, it should include some information on their combinability with the items in the stem lexicon (e.g. their word-class affiliations) together with their grammatical functions. The stem lexicon will be based on the wordlist of the BT and apart from the stems themselves will include some additional grammatical and morphological information about each item.

The creation of the stem lexicon poses several difficulties. First, a wordlist has to be extracted from the dictionary and for that purpose, the headwords have to be tagged unambiguously (this has already been partially done for the preliminary dictionary application). Second, the additional grammatical and morphological information needs to be extracted and therefore the appropriate elements in the dictionary's microstructures have to be isolated and tagged. This additional information is of two types.

The morphological information should constitute what Osolsobě termed an intersegment, i.e. the part of a stem that changes according to which ending is appended to the stem. For example, the nominal paradigm of *fugol* ("bird") syncopates the pre-final "o" in all cases except Nom. and Acc., e.g. Gen. *fugles*. This obviously depends on our understanding of the stem and can greatly affect the overall efficiency of the resulting analyzer—we could also consider *fug* to be a stem

---

[6] "The Text Encoding Initiative (TEI) Guidelines are an international and interdisciplinary standard that enables libraries, museums, publishers, and individual scholars to represent a variety of literary and linguistic texts for online research, teaching, and preservation."

[7] The word-formation analyzer would be particularly interesting as well as quite demanding to design, since there is no comprehensive monograph mapping Old English word-formation as yet.

[8] Another lexicon of derivational prefixes may be considered at a later stage.

and the endings would then be *-ol* or *-les*. This is obviously nonsense in terms of Old English morphology—we consider it here just from the point of view of the analyzer's operation. Such treatment of morphology, however, might greatly increase the level of ambiguity in the stem lexicon and increase the diversity in the lexicon of endings. Moreover, it does not seem wise to diverge in our divisions of forms into lexicons from the actual Old English morphology. Although it might appear fruitful at this stage of morphological analysis (we would not need to include the intersegmental information at all), it could prove fatal if we tried to extend our research into the field of word-formation.

The intersegments will have to be identified manually, using the recurrent patterns of stem-final syllables. The rest of the additional morphological information should consist of variant spellings that can be extracted from the dictionary.

The grammatical part of the additional information should contain mainly the word-class and gender affiliation, possibly the verb-class and "stem-type" affiliations. As this information will serve to pair the stems with endings, the more detailed information provided for each stem, the better the chance to get a "permissible" result form.[9] In other words, by adding more relevant information to the stem lexicon we try to follow the structural workings of an inflectional language. If the analyzer is found to return a large number of non-permissible forms, we may need to add more detailed information to improve pairing of items between the lexicons.[10]

Another difficulty in creating the stem lexicon arises from the fact that the word-list consists of lemmata, rather than bare stems. In order to transform the lemmata into stems, we need a simple version of a stemmer tool that would strip the lemmata of their "lemma-forming"[11] endings. This might prove tricky because some of these endings might match strings of characters at the end of inflexible words or even the final part of different endings. For this reason, the grammatical information should be extracted from the dictionary prior to this process, so that the cropping can be run only on subsets of the lexicon, reducing thus the chance of wrong matches (so the verbal endings will be removed only from verbal stems, etc.). It is expected that this process will need a great amount of manual checking and correcting.

The problem of introflection and a general variability of the stem has already been noted and an automatic analyzer will have to deal with it. The problem can be again divided into two parts: the grammatical variation of a root vowel (ablaut and umlaut) and the non-grammatical variations of the root vowel. The grammatical variation can be dealt with in two ways. The information about the permissible vowels can be either included with the stem (e.g. as a part of the intersegment information), or it can be stored separately as a kind of filter consisting of probable variations based on grammatical information supplied with the stem and its morphology (e.g. vowel harmony, ablaut in strong verbs, vowel shortening before consonant clusters, etc.). Non-grammatical variations, i.e. variations not associated with grammatical function, include

[9] The permissible result forms are such forms that conform to the standard descriptions of Old English morphology and grammar we are working with (like Campbell's), whether they be attested or unattested ("missing") forms.

[10] It is worth noting here that Osolsobě's model works with exactly paired items: it is possible to infer all and only the permissible combinations of stems and endings from her lexicon. We choose a slightly different approach (our lexicons would generate all the probable forms including many non-permissible ones), because Osolsobě's method requires manually supplying the pairing information for each stem (although she facilitates the process by introducing a great number of permissible paradigm patterns). Moreover, compared to Czech, Old English makes a proportionally greater use of introflection and smaller use of endings, so that the number of stem+ending combinations should be much smaller. Last but not least, it might be of theoretical interest to see how the number of combination decreases with the amount of grammatical information we supply for stems.

[11] That is endings used to form a lemma, e.g. ending of nom. sg. for nouns or 1st person sg. indicative present for verbs.

dialectal, diachronic or orthographic variations. These will be partly covered as stem variants, but the more regular variations should also be fed into the above-mentioned filters (e.g. dialectal or scribal variation in vowels: WS[12] *eald* / Angl.[13] *ald*, WS *sprecan* / Kt.[14] *spreocan*, WS *dehter* / Nth. *dæhter*; diachronic variation: early WS *cneoht* / late WS *cniht*; etc.). The filters could then be applied either in case the grammatical information matches any possible grammatical root vowel variation or in case no corresponding stem can be identified in the stem lexicon.

The course of the analyzer's operation

There seem to be two basic possibilities for the actual course of the analyzer's operation from the user's input to an identified lemma and its grammatical information on the output. Based on the above-mentioned principles, we can follow the procedure proposed by Osolsobě and analyze the user's input character by character from right to left. First we would identify the ending in the lexicon of endings and succeeding or failing in that we would identify the appropriate intersegment and stem. This course of operation might seem the most logical. However, it has been shown as not quite practical by Sedláček & Smrž. First, it complicates the use of the above-mentioned filters—at which character should the analyzer start looping for the possible variations?—, especially those that are grammatically conditioned, because the analyzer is logically unaware of any potential grammatical information the output lemma might provide before its operation is finished. The method Sedláček & Smrž come up with does not only seem to provide solution to these problems, but also offers a faster operation, at least in case of Czech. Instead of character by character comparison of the input with the two basic lexicons, the input is compared with a third lexicon consisting of all possible combinations of the two basic lexicons. This third lexicon is prepared during the analyzer's development stage and although the product is more bulky in terms of the amount of its data, it is simpler and faster in operation. The two above-mentioned problems are solved, because the filters (both conditioned and unconditioned) are used during the generation of the third lexicon already.

Conclusion

It is clear that the results of the analyzer as suggested above will be only partially correct. For example, combining all nominal endings with all nominal stems will obviously generate a large number of not only unattested, but also wholly impossible or non-permissible forms (combinations of endings and stems belonging to different paradigms, etc). However, the connection of the analyzer with the dictionary suggests that the main use for the analyzer will be to identify possible lemmata in the dictionary (i.e. dictionary entries) from the user input and perhaps to supply additional grammatical information about them as well. With this purpose in mind it seems better to generate some nonsensical data that will usually not match with the dictionary wordlist anyway, rather than possibly omit correct matches. One way of redressing the problem of nonsensical data could be the inclusion of a list of types from an Old English Corpora so that the analyzer could notify the user that a particular generated form is unattested.[15]

With a morphological analyzer and with an appropriate tagging, the digitized *Anglo-Saxon Dictionary* might be transformed into a tool whose possibilities would greatly surpass its printed version, an instrument from which not only English medievalists, but all scholars

[12] West Saxon.

[13] Anglian.

[14] Kentish.

[15] Here again we may benefit from not following Osolsobě's approach strictly, because the investigation of the generated permissible but unattested forms might prove fruitful. Many of these forms could in fact have been used, but were either never recorded or their records were lost to us. Some of the forms might also point to cases where analogy would have produced them in the end, if other processes had not intervened, etc.

interested in Old English and the development of English and Germanic languages could greatly benefit. We are optimistic about the future of the project for three reasons in particular: a similar approach has now proved efficient in the case of Modern Czech; a large part of the work has already been done by digitizing the BT; and, last but not least, the project has now received support from the Charles University Grant Agency.

Bibliography

Bosworth, J.; Toller, T. N. (1898-1921). *An Anglo-Saxon dictionary, based on the manuscript collections of the late Joseph Bosworth*. Oxford: Oxford University Press.

Campbell, A. (1959). *Old English Grammar*. Oxford: Oxford University Press.

Crist, S. (2001). *Germanic Lexicon Project* [on line]. *http://lexicon.ff.cuni.cz/about/aa_project_goals.html* [Access date: 29 March 2008].

Osolsobě, K. (1996). *Algoritmický popis české formální morfologie a strojový slovník češtiny*. (Unpublished dissertation). Brno: Masaryk University.

Sedláček, R.; Smrž, P. (2001). "Automatic Processing of Czech Inflectional and Derivative Morphology". FI MU Report Series June 2001. 2-13.