The HeliPaD: a parsed corpus of Old Saxon

George Walkden
University of Manchester

This short note introduces the HeliPaD, a new parsed corpus of Old Saxon (Old Low German). It is annotated according to the standards of the Penn Corpora of Historical English, enriched with lemmatization and additional morphological attributes as well as textual and metrical annotation. This note provides an overview of its main features and compares it to existing resources such as the *Deutsch Diachron Digital* version of the Old Saxon *Heliand* as part of the *Referenzkorpus Altdeutsch*. It closes with a roadmap for planned future expansions.

Keywords: Old Saxon, Low German, parsed corpus, historical corpus

1. A new corpus for Old Saxon

The aim of this paper is to introduce the HeliPaD, a new parsed corpus of Old Saxon (Old Low German). It is annotated according to the standards of the Penn Corpora of Historical English, enriched with lemmatization and additional morphological attributes as well as textual and metrical annotation. The rest of Section 1 gives an overview of the corpus and the language, and Section 2 compares it to existing resources such as the *Deutsch Diachron Digital* version of the Old Saxon *Heliand*. Section 3 goes into more detail on the annotation of the corpus, and Section 4 summarizes and concludes.

1.1 The Old Saxon language

Old Saxon (also known as Old Low German) is a West Germanic language that was spoken in the area of what is now northern Germany before 1100 AD. It is usually thought to be the ancestor of the Middle Low German language, though the extent to which there is continuity between the language represented in the extant Old Saxon texts and that represented in Middle Low German texts is a matter of debate. Rauch (1992) and Cathey (2000) provide accessible scholarly introductions to the language, and Cordes & Holthausen (1973) and Gallée & Tiefenbach (1993) are the standard Old Saxon grammars. Old Saxon is transmitted in two main texts: the *Heliand* (which represents the vast majority of attested Old Saxon), and a verse translation of *Genesis*. In addition, there are a number of shorter texts of no more than a few paragraphs each, as well as a number of glosses.

The *Heliand* is a gospel harmony written in alliterative verse, and a very loose translation (or recomposition) of Tatian's *Diatessaron* (Grein 1869).[1] The original text is presumed to date from the first half of the 9th century. In total, 5,983 lines have been preserved, in six manuscripts: C (Cotton), M (Monacensis), S (Straubing), V (Vatican), P (Prague), and L (Leipzig). The S, V, P and L manuscripts are extremely limited in extent, and none of them contains a continuous stretch of more than a hundred lines. The M and C manuscripts are the main witnesses to the text. While the M manuscript contains a number of gaps, the C manuscript (Cotton Caligula A VII, British Library) is complete up to line 5,968. The text is divided into 71 sections, called fitts.

There exist two editions of the *Heliand* with present relevance: Sievers (1878), a broadly diplomatic edition of manuscripts C and M, and Behaghel (1902 and nine subsequent editions, e.g. Behaghel & Taeger 1984), the standard critical edition. Dewey (2009) is a modern English translation with textual notes.

1.2 Introducing the HeliPaD

The HeliPaD (*Heliand* Parsed Database, Walkden 2015a) is a richly annotated version of the *Heliand*, annotated according to the standard used by the Penn Corpora of

Historical English (e.g. YCOE, Pintzuk et al. 2003; PPCME2, Kroch & Taylor 2000) and related parsed corpora (e.g. IcePaHC, Wallenberg et al. 2011, and see Rögnvaldsson et al. 2012, for historical Icelandic; MCVF, see Martineau 2008, for historical French; Tycho Brahe Corpus, Galves & Faria 2010, for historical Portuguese) and designed to be queried using CorpusSearch 2 (Randall 2005-7). The total size of the corpus is 46,067 words (not including punctuation and code). It is available for download at http://www.chlg.ac.uk/helipad/, and has been released under a Creative Commons CC-BY 4.0 license.

As source text for the HeliPaD, the text of the C manuscript as reproduced in Sievers (1878) was used. This choice calls for some comment, as it differs from that of other online Old Saxon resources: both the TITUS version of the *Heliand* (Gippert 2003) and the *Referenzkorpus Altdeutsch* version are based on Behaghel & Taeger (1984). While the original intention of the *Referenzkorpus Altdeutsch* team was to include a close transcription of the manuscript alongside the edited version (Linde & Mittmann 2013: 236), this eventually proved unfeasible for reasons of cost (Donhauser 2015: 37). The essential problem with the Behaghel edition is twofold. First, it is a critical edition that conflates several manuscript sources, some of which are demonstrably distinct in their linguistic features. This distinctness is found even at the syntactic level: for instance, pronominal subjects are often present in the M manuscript while they are absent in the C manuscript (Walkden 2014a: 192-193). Secondly, the Behaghel edition contains a very large number of emendations. It is an edition created for the purposes of the general reader and literary scholar rather than for the serious linguist, and in a recent comparative study of the *Heliand* manuscripts Price (2010) concludes that "despite being regarded as the standard resource in Old Saxon studies, Behaghel's rendition of the *Heliand* is highly unreliable" (2010: 54-55).

Sievers' (1878) parallel edition, on the other hand, presents versions of the two main manuscripts – C and M – on facing pages, and Price notes that "Sievers' transcription is much more apt to represent the text true to the manuscripts" (2010: 56), though it does contain a number of misprints as well as emendations, so that it cannot be taken as a faithful transcription (Price 2010: 56–58). Though the HeliPaD has faithfulness to the source manuscript(s) as its primary goal, a new transcription would have been too time-consuming; in the future I hope to be able to carry out a full

collation of Sievers' edition with Cotton Caligula A VII. A final major advantage of the Sievers (1878) edition is that it is unquestionably in the public domain.

The C manuscript was chosen primarily because it is the most complete. Manuscript M, despite being the preferred source for Behaghel's edition, contains a number of gaps, including incomplete clauses. Since the HeliPaD is first and foremost a tool for syntactic research, the C manuscript was preferable. C may also be closer to the archetype than M (Behaghel & Taeger 1984: xvii-xviii). The HeliPaD contains all 5,968 lines of Sievers' edition of C.

2. Comparison with existing resources

Until recently, the only searchable version of the *Heliand* available was part of the *Thesaurus Indogermanischer Text- und Sprachmaterialien* (TITUS) (Gippert 2003), which can be queried online only by word-form. However, in addition to the present resource and its predecessor, the HeliCoPTeR (Walkden 2011), a version of the *Heliand* has also been made available recently as part of the *Deutsch Diachron Digital* (DDD) reference corpus of Old German (*Referenzkorpus Altdeutsch*), with rich multilayered annotation (Linde & Mittmann 2013; Donhauser 2015). This section outlines the similarities and differences between these two resources.

One major difference has already been mentioned: the HeliPaD is based on Sievers' (1878) edition of the C manuscript, while the DDD version is based on the critical edition by Behaghel & Taeger (1984). This means that users who are particularly interested in the forms found in a particular manuscript will be better served by the HeliPaD at present – though, due to the drawbacks of all existing editions discussed in Price (2010), such users are really in need of either a facsimile edition or a brand new diplomatic edition. On the other hand, following the Behaghel & Taeger (1984) edition, the DDD *Heliand* contains vowel length markings, which can be considered a form of phonological annotation given that they are not present in the manuscript, albeit annotation that is not carried out according to explicitly stated principles.

Morphologically, both the HeliPaD and the DDD *Heliand* contain a great deal of overlapping information. Both are lemmatized and POS-tagged, and marked for the attributes person, number and case. In addition, the DDD *Heliand* contains gender as an attribute, and is also annotated for inflectional class (e.g. strong verb class) – though the latter is in principle fully retrievable on the basis of the lemma in the HeliPaD. For morphological purposes, then, if gender or inflectional class is key, the DDD *Heliand* is the resource to use; otherwise, the HeliPaD will serve equally well.

Both resources also contain textual and metrical annotation: fitts and line/verse numbers are found in both. In the HeliPaD the position of the caesura is also marked, as are the manuscript page breaks and pages from Sievers (1878), while the DDD *Heliand* is annotated for alliteration, and has also been aligned to the Old High German *Diatessaron* and Otfrid's *Evangelienbuch* (see Price 2015). Here, then, which resource to use depends on the specific features one is interested in.

It is in the domain of syntax where the HeliPaD is likely to be of most use. The DDD *Heliand* is chunk-parsed into clause-level units which are annotated for function (e.g. adverbial concessive clause). The HeliPaD, on the other hand, is fully parsed: the structure of noun phrases, grammatical relations, long-distance dependencies, and more are all represented in the annotation, as discussed in the following section – a substantial advantage for the historical syntactician.

In summary, both resources are useful for research into Old Saxon, and they are to some extent complementary. Which to use depends partly on the linguistic level under investigation – syntax, morphology, (supra)segmental phonology – and partly on whether a (mostly) diplomatic edition is to be preferred.

3. Annotation

The HeliPaD is annotated on a number of levels. Figure 1 shows the structure of a typical token, as visualized in CorpusDraw (Randall 2005-7), which converts the labeled bracketing representations of the file into a tree representation. The sentence can be translated into English as "They were dear to God, worthy of the work".

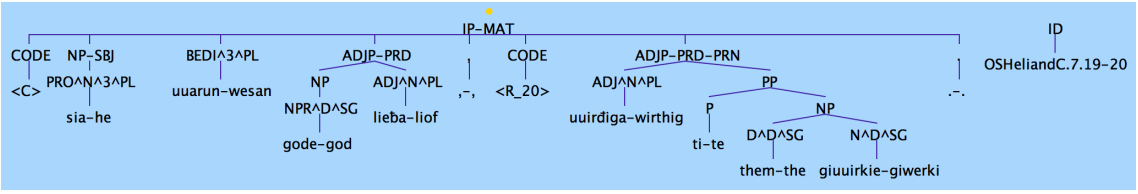

Figure 1: A token from the HeliPaD

A token in the Penn Corpora and HeliPaD is, broadly speaking, a main verb and everything that belongs with it.[2] In many cases, it will be a "sentence", in pretheoretical terms. The main exception is when two independent clauses with finite verbs are conjoined, in which case these are treated as separate tokens. The token is enclosed in brackets, and consists of a parse followed by a token ID, each of which is itself enclosed in brackets. The ID takes the form *OSHeliandC.foo.bar*, where *foo* is simply a sequential number starting at 1 and *bar* is the range of lines spanned by the token. For instance, *OSHeliandC.265.502-503* is the ID for a token that starts on line 502 and ends on line 503 and is the 265th token in total.

The corpus character encoding is UTF-8, and contains certain special characters such as barred b (ƀ) and d (đ). There is thus no need for specifically orthographic annotation. The corpus does, however, contain substantial amounts of metatextual information, including manuscript page, Sievers edition page, fitt, line, caesura (half-line break), and other comments (e.g. in case of textual problems). Figure 1 contains a line break, the start of line 20, marked as <R_20> in angle brackets and POS-tagged as CODE. There is also a caesura at the beginning of the token, marked as <C>.

A significant difference between the Penn Corpora of Historical English and the HeliPaD (and a property that the HeliPaD shares with the IcePaHC) is that the HeliPaD is lemmatized. The lemma is given after the word form and separated by a hyphen: thus, for the second person singular present indicative of the verb *wesan* 'to be', what is found in the corpus is *bist-wesan*. Lemmas are based in form on the freely available Old Saxon dictionary and concordance by Köbler (2000), though Köbler's length markings are not reproduced in the corpus.

The most in-depth annotation in the HeliPaD is for morphological and syntactic properties, and here the HeliPaD most closely follows the York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE; Taylor et al. 2003). It contains POS-tagging using a set of 88 tags, augmented with the morphological attributes case,

number and person, and parsing following the YCOE's guidelines. These are the subject of the following subsections.

Since the corpus is small, all annotation was carried out using scripts written for the purpose and then hand-corrected. For the bulk of the syntactic parsing, the program Annotald (Beck, Ecay & Ingason 2015), which provides a GUI for annotation in the Penn Treebank format, was used.

3.1 Morphology

The part-of-speech annotation of the Penn Historical Corpora is based on that of the Penn Treebank (Santorini 1990). This tagset was designed with English, a predominantly analytic language, in mind. Old Saxon, like Old English, is further towards the synthetic end of the continuum, with a productive system of grammatical case as well as robust agreement for person and number. There is therefore information in the text(s) that would not be captured by straightforwardly transferring the Penn tagset. It is desirable to capture this information in the morphological annotation, and to do so without losing generalizations, i.e. paradigmatic relations between tags.

The YCOE (Taylor et al. 2003) extends the tagging of the other Penn corpora by adding case as an attribute separated from the tag by a caret: N^D, for instance, is a noun in the dative case. This attribute-based system is also adopted in the IcePaHC (Wallenberg et al. 2011). The HeliPaD takes this approach further by adding two further attributes: person and number. Nominal tags (nouns, attributive adjectives, inflected participles, quantifiers, determiners, numerals, pronouns) are specified for case: N(ominative), A(ccusative), D(ative), G(enitive) or I(nstrumental). Finite verbal tags are specified for person (1, 2, or 3), as are pronouns. Both nominal and finite verbal tags are specified for number: SG (singular), DU (dual) or PL (plural). Pronouns are the only tag to bear all three attributes, which occur in the order case > person > number. For example, the form *ik*, a first person singular nominative pronoun, would be lemmatized as *ik* and tagged as PRO^N^1^SG. In the corpus file, this would be displayed as (PRO^N^1^SG *ik*-*ik*).

This simple extension of the Penn Historical Corpora tagging scheme enables a much wider range of morphological information to be represented. With minimal modification, the attribute-vector format of the fine-grained tags also allows for the corpus to serve as input to a tagger such as RFTagger (Schmid & Laws 2008), which uses decision trees to estimate attribute probabilities and outperforms n-gram-based taggers for languages with rich morphology. This will come in useful for the annotation of further Old Saxon texts.

3.2 Syntax

The parsing of the HeliPaD follows that of the YCOE closely. In the online manual at http://www.chlg.ac.uk/helipad/ there is a full guide to the syntactic annotation principles, as well as a summary page listing the points where the HeliPaD's policy diverges from that of the YCOE. Following the Penn principles, the HeliPaD annotation is not designed to represent the last word on the syntactic structure of Old Saxon, but rather to ease retrieval of particular types of phrase by the analyst. The emphasis is on clearly documented, replicable annotation principles, and on minimizing the number of cases where a subjective judgment call would have to be made. A case in point is the absence of a VP node, a feature which the HeliPaD shares with all Penn historical corpora. Due to the flexibility of clausal word order and the difficulty of deciding where a VP would begin and end, this level of structure is simply not represented in the annotation: instead the IP node dominates all clause-internal material (see Figure 1). This does not imply that the annotators believe, or are advancing the hypothesis, that these languages have no VP – though this claim has been made (see e.g. Faarlund 1990 and the discussion in Rögnvaldsson 1995 for Old Norse-Icelandic). On controversial points, the annotation itself remains silent. On the other hand, all annotation is driven by theory (see Xiao 2009: 995-996 for lucid discussion), and I can make no claim to complete objectivity or accuracy for the HeliPaD's parsing. For instance, the division between main clauses (IP-MAT) and subordinate clauses (IP-SUB) is based partly on the implicit analysis presented in the sentential punctuation of Sievers (1878), since many adverbial subordinators are homophonous with independent adverbs. Yet the

manuscript contains little punctuation: only a *punctus* is regularly used, and Sievers (1878: xii, xiv) refers to its placement as completely arbitrary ("vollkommen willkürlich"). Sievers therefore punctuates the Heliand according to nineteenth-century German norms; see Price (2010: 59) for critique. The division into main clauses and subordinate clauses may not be robust, therefore, and linguists interested in this question should not take the annotation as God's truth (cf. Rissanen 1989). Somers & Dubenion-Smith (2011) present a cautious and philologically-informed approach to the issue of clausal status in Old Saxon.

The existence of a parsed corpus should, however, greatly facilitate the study of Old Saxon syntax. This has been a neglected topic in the last century, especially when compared to the syntax of its better-attested Insular sister, Old English. The position of the finite verb has attracted some attention (Erickson 1997, Dewey 2006, Hinterhölzl & Petrova 2009, Somers & Dubenion-Smith 2014, Walkden 2014a, 2015b and references cited there); there is also published work on auxiliaries and auxiliary selection (Arnett 1997, Watts 2001), predicative adjective agreement (Hock 2009), reflexivization (Sapp 2010), negation (Breitbarth 2014), subject omission (Walkden 2014a), object position (Walkden 2014b), and case and grammatical relations (Dewey & Arnett 2015). These topics barely scratch the surface of what could be investigated; moreover, due to the lack of corpus resources until recently, few of the aforementioned works present detailed quantitative information.

A problem often adduced in the context of Old Saxon syntax is that the main textual witnesses, the *Heliand* and *Genesis*, are written in alliterative verse that is constrained to comply with very specific metrical requirements (Sievers 1875; Suzuki 2004). As such, any conclusions drawn on the basis of these documents may be unlikely to generalize to Old Saxon as it was actually used by most of its speakers most of the time. To some extent, this problem is inescapable, and is therefore relevant only insofar as conclusions about Old Saxon word order must be presented with appropriate hedges and caution. On the other hand, the advent of modern statistical techniques for multivariate analysis means that in some cases it is possible to isolate, and quantify, the contribution of prosodic factors to word order alternations: Walkden (2014b), for instance, uses logistic regression analysis to estimate the effect of heavy prosodic weight in favouring a late position for objects. Since such techniques require a large

amount of data in order to yield reliable results, which until recently had to be hand-collected, the availability of a parsed corpus should make it substantially easier to carry out studies of this sort. Moreover, for certain phenomena the metrical "confound" could turn out to be a positive boon. Pintzuk & Kroch (1989), for instance, are able to make the case for the Old English poem *Beowulf* that PP extraposition and heavy NP shift are two distinct processes, since postverbal NPs occur in a different half-line significantly more frequently than do postverbal PPs; this type of study is impossible using prose sources. It is now possible to carry out a parallel study on Old Saxon using the HeliPaD, and much more besides.

4. Conclusion and future plans

This note has introduced a new parsed corpus of Old Saxon, which I hope will be useful for research into the history of Low German and the comparative philology of early Germanic. The corpus is free to use, and is released under a CC-BY 4.0 license. Rögnvaldsson et al. (2012: 1982) outline ten desiderata for user freedom. According to these, a corpus should be:

i. Available for download as raw data;
ii. Comprehensively documented online;
iii. Available without registration;
iv. Developed using free/open source tools;
v. Developed in an open online version control repository;
vi. Released in numbered versions for version control;
vii. Modifiable by users without special permission;
viii. Free of cost to academia;
ix. Free of cost to commercial users;
x. Released under a standard free license (e.g. GPL, LGPL, CC).

The HeliPaD meets nine out of ten of these criteria: the development process did not involve an open version control repository (desideratum 5), largely because making

incomplete versions of a single, substantial text available did not seem to make much sense. The current version is 0.9: some errors undoubtedly remain, and corrections are welcomed.

In the future I intend to expand the corpus in at least two ways. First, the remaining Old Saxon textual material – *Genesis*, plus other minor texts in Köbler (1986) – will be annotated to the same standard and incorporated into the corpus. Secondly, and subsequently, the other manuscripts of the *Heliand* will be annotated and aligned with the C manuscript; this will be a substantial task only for the M manuscript, for which a facsimile has already been made available online by the Bayerische Staatsbibliothek. A third step will be the collation of the Sievers (1878) edition with the actual C manuscript, in essence preparing a new transcription/edition, though this is a far more substantial task if done properly. In the meantime, I would be happy to receive comments, criticisms and revisions, and even happier to see the corpus put to good use in research on Old Saxon.

Acknowledgements

I am indebted to Anne Breitbarth, Tonya Kim Dewey and Svetlana Petrova for discussion of the Old Saxon material in the planning stage, and to Aaron Ecay, Anton Karl Ingason, Tony Kroch, Caitlin Light, Susan Pintzuk, Ann Taylor, Joel Wallenberg, and David Willis for their practical guidance on how to get things done. Special thanks go to Sheila Watts, who carried out part of the morphological annotation as well as being a constant source of useful advice. I am also grateful to Gavin Brookes and two anonymous IJCL reviewers for their comments on an earlier draft of this paper.

Notes

1. The *Diatessaron* is a conflation of the four Gospel stories into one narrative, and was composed by Tatian in either Syriac or Greek during the 2nd century CE. There exists also an Old High German translation of the *Diatessaron*, which is more commonly known as the Old High German *Tatian*. Whether (a manuscript of) the Latin version was the primary source text

for the *Heliand*, or rather the Old High German version, as proposed by Baesecke (1948), is debated.

2. Note that this use of the term 'token' is substantially different from its use in traditional corpus linguistics, in which the term is used in opposition to 'type', to denote an item found in a corpus that can be categorized as belonging to a particular type. See McEnery & Wilson (2001: 82) for discussion in that context.

References

Arnett, C. (1997). Perfect auxiliary selection in the Old Saxon *Heliand*. *American Journal of Germanic Linguistics and Literatures*, 9, 23-72.

Baesecke, G. (1948). Fulda und die altsächsischen Bibelepen. *Niederdeutsche Mitteilungen*, 4, 5-43.

Beck, J., Ecay, A., & Ingason, A. K. (2015). *Annotald*. Version 1.3.7. Retrieved from https://github.com/Annotald/ (last accessed January 2016).

Behaghel, O. (ed.) (1902). *Der Heliand und die altsächsische Genesis* (1st ed.) Gießen, Germany: J. Ricker'sche Buchhandlung.

Behaghel, O., & Taeger, B. (eds.) (1984). *Heliand und Genesis* (9th ed.) Tübingen, Germany: Max Niemeyer.

Breitbarth, A. (2014). *The History of Low German Negation*. Oxford, UK: Oxford University Press.

Cathey, J. E. (2000). *Old Saxon*. Munich, Germany: Lincom Europa.

Cordes, G., & Holthausen, F. (1973). *Altniederdeutsches Elementarbuch*. Heidelberg, Germany: Carl Winter.

Dewey, T. K. (2006). *The origins and development of Germanic V2* (Unpublished doctoral dissertation). University of California, Berkeley, CA.

Dewey, T. K. (2009). *An Annotated English Translation of the Old Saxon* Heliand*: A Ninth-Century Biblical Paraphrase in the Germanic Epic Style*. New York, USA: Edwin Mellen Press.

Dewey, T. K., & Arnett, C. (2015). Motion Verbs in Old Saxon with the Oblique Subject Construction: A Semantic Analysis. *Beiträge zur Geschichte der Deutschen Sprache und Literatur*, 137, 183-220.

Donhauser, K. (2015). *Das Referenzkorpus Altdeutsch: Das Konzept, die Realisierung und die neuen Möglichkeiten*. In J. Gippert & R. Gehrke (Eds.), *Historical Corpora: Challenges and Perspectives* (pp. 35-49). Tübingen, Germany: Narr Verlag.

Erickson, J. (1997). Some observations on word order in Old Saxon. In C. Dürscheid, K H. Ramers, & M. Schwarz (eds.), *Sprache im Fokus: Festschrift für Heinz Vater zum 65. Geburtstag* (pp. 95-105). Tübingen, Germany: Max Niemeyer.

Faarlund, J.-T. (1990). *Syntactic change: towards a theory of historical syntax*. Berlin, Germany: Mouton de Gruyter.

Gallée, J. H., & Tiefenbach, I. (1993). *Altsächsische Grammatik*. Tübingen, Germany: Max Niemeyer.

Galves, C., & Faria, P. (2010). *Tycho Brahe Parsed Corpus of Historical Portuguese*. Retrieved from http://www.tycho.iel.unicamp.br/~tycho/corpus/en/index.html (last accessed January 2016).

Gippert, J. (2003). *TITUS: Heliand*. Retrieved from http://titus.uni-frankfurt.de/texte/etcs/germ/asachs/heliand/helia.htm (last accessed January 2016).

Grein, C. W. M. (1869). *Die Quellen des Heliand. Nebst einem Anhang: Tatians Evangelienharmonien herausgegeben nach dem Codex Cassellanus*. Kassel, Germany: T. Kay.

Hinterhölzl, R., & Petrova, S. (2009). From V1 to V2 in West Germanic. *Lingua*, 120, 315-328.

Hock, H. H. (2009). Default, animacy, avoidance. In V. Bubenik, J. Hewson, & S. Rose (Eds.), *Grammatical Change in Indo-European Languages* (pp. 29-42). Amsterdam, Netherlands: John Benjamins.

Köbler, G. (1986). *Sammlung aller altsächsischer Texte*. Gießen, Germany: Arbeiten zur Rechts- und Sprachwissenschaft Verlag.

Köbler, G. (2000). *Altsächsisches Wörterbuch* (3rd ed.) Retrieved from http://www.koeblergerhard.de/aswbhinw.html (last accessed January 2016).

Kroch, A., & Taylor, A. (2000). *Penn-Helsinki Parsed Corpus of Middle English* (2nd ed.) Retrieved from http://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-3/ (last accessed January 2016).

Linde, S., & Mittmann, R. (2013). Old German reference corpus: Digitizing the knowledge of the 19th century. In P. Bennett, M. Durrell, S. Scheible & R. J. Whitt (Eds.), *New Methods in Historical Corpora*, 235-246. Tübingen, Germany: Narr Verlag.

Martineau, F. (2008). Un corpus pour l'analyse de la variation et du changement linguistique. *Corpus*, 7, 135-155.

McEnery, A., & Wilson, A. (2001). *Corpus linguistics: an introduction* (2nd ed.) Cambridge, UK: Cambridge University Press.

Pintzuk, S., & Kroch, A. (1989). The rightward movement of complements and adjuncts in the Old English of *Beowulf*. *Language Variation and Change*, 1, 115-143.

Price, T. B. (2010). *The Old Saxon Leipzig Heliand manuscript fragment (MS L): New evidence concerning Luther, the poet, and Ottonian heritage* (Unpublished doctoral dissertation). University of California, Berkeley, CA.

Price, T. B. (2015). Multi-faceted alignment: Toward automatic detection of textual similarity in Gospel-derived texts. In J. Gippert & R. Gehrke (Eds.), *Historical Corpora: Challenges and Perspectives* (pp. 77-89). Tübingen, Germany: Narr Verlag.

Randall, B. (2005–7). *CorpusSearch 2*. Retrieved from http://corpussearch.sourceforge.net/ (last accessed January 2016).

Rauch, I. (1992). *The Old Saxon language: grammar, epic narrative, linguistic interference*. Frankfurt, Germany: Peter Lang.

Rissanen, M. (1989). Three problems connected with the use of diachronic corpora. *ICAME Journal*, 13, 16-19.

Rögnvaldsson, E. (1995). Old Icelandic: A non-configurational language? *North-Western European Language Evolution*, 26, 3-29.

Rögnvaldsson, E, Ingason, A. K., Sigurðsson, E. F., & Wallenberg, J. (2012). The Icelandic Parsed Historical Corpus (IcePaHC). In N. Calzolari, K. Choukri, T. Declerck, M. U. Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & Stelios Piperidis (Eds.), *Proceedings of LREC, Istanbul 2012* (pp. 1977-1984). Istanbul, Turkey: European Language Resources Association.

Santorini, B. (1990). *Part-of-speech tagging guidelines for the Penn Treebank Project* (3rd revision). Technical Report, University of Pennsylvania Department of Computer & Information Science.

Sapp, C. (2010). The Reflexive Possessive *sîn* in Old Saxon. *Beiträge zur Geschichte der Deutschen Sprache und Literatur*, 132, 329-342.

Schmid, H., & Laws, F. (2008). Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-Grained POS Tagging. In D. Scott & H. Uszkoreit (Eds.), *Proceedings of the 22nd International Conference on Computational Linguistics*, vol. 1 (pp. 777-784). Stroudsburg, PA, USA: Association for Computational Linguistics.

Sievers, E. (1875). *Der Heliand und die angelsächsische Genesis*. Halle, Germany: Max Niemeyer.

Sievers, E. (ed.) (1878). *Heliand*. Halle, Germany: Verlag der Buchhandlung des Waisenhauses.

Somers, K., & Dubenion-Smith, S. (2011). Disambiguating clausal status in the Old Saxon *Hêliand*. Ms., QMUL.

Somers, K., & Dubenion-Smith, S. (2014). The intersection between syntax and metre in the Old Saxon *Hêliand*. *Amsterdamer Beiträge zur älteren Germanistik*, 72, 83-134.

Suzuki, Seiichi. 2004. *The Metre of Old Saxon Poetry: The Remaking of Alliterative Tradition*. Cambridge, UK: Brewer.

Taylor, A., Warner, A., Pintzuk, S., & Beths, F. (2003). *York–Toronto–Helsinki Parsed Corpus of Old English Prose*. Retrieved from http://www-users.york.ac.uk/~lang22/YCOE/YcoeHome.htm (last accessed January 2016).

Walkden, G. (2011). *HeliCoPTER: Heliand Corpus, a Partially Tagged Excel Resource*. Version 1.0. Retrieved from http://personalpages.manchester.ac.uk/staff/george.walkden/helicopter/ (last accessed January 2016).

Walkden, G. (2014a). *Syntactic Reconstruction and Proto-Germanic*. Oxford, UK: Oxford University Press.

Walkden, G. (2014b). Object position and Heavy NP Shift in Old Saxon and beyond. In K. Bech & K. M. Eide (Eds.), *Information Structure and Word Order Change in Germanic and Romance Languages* (pp. 313-340). Amsterdam, Netherlands: John Benjamins.

Walkden, G. (2015a). *HeliPaD: the Heliand Parsed Database*. Version 0.9. Retrieved from http://www.chlg.ac.uk/helipad/ (last accessed January 2016).

Walkden, G. (2015b). Verb-third in early West Germanic: a comparative perspective. In T. Biberauer & G. Walkden (Eds.), *Syntax over time: lexical, morphological, and information-structural interactions* (pp. 236-248). Oxford, UK: Oxford University Press.

Wallenberg, J. C., Ingason, A. K., Sigurðsson, E. F., & Rögnvaldsson, E. (2011). *Icelandic Parsed Historical Corpus (IcePaHC).* Version 0.9. Retrieved from http://www.linguist.is/icelandic_treebank/ (last accessed January 2016).

Watts, S. (2001). How to become an auxiliary: progressive and perfect in Old Saxon. In S. Watts, J. West, & H.-J. Solms (Eds.), *Zur Verbmorphologie germanischer Sprachen* (pp. 117-135). Berlin, Germany: Mouton de Gruyter.

Xiao, R. (2009). Theory-driven corpus research: Using corpora to inform aspect theory. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook* (pp. 987-1008). Berlin, Germany: Mouton de Gruyter.

Author's address

George Walkden
School of Arts, Languages and Cultures
University of Manchester
Oxford Road
Manchester M13 9PL
UK

george.walkden@manchester.ac.uk