The Design and Use of a Latin Dependency Treebank

David Bamman and Gregory Crane
The Perseus Project, Tufts University

{david.bamman, gregory.crane}@tufts.edu

1 Introduction

While much of the research and labor in treebanks has focused on modern languages, recent scholarship has also seen the rise of treebanks for historical languages as well, such as Middle English (Kroch and Taylor [15]), Early Modern English (Kroch et al. [16]), Old English (Taylor et al. [28]), Early New High German (Demske et al. [11]) and Medieval Portuguese (Rocio et al. [27]). Like their modern counterparts, these historical treebanks serve two distinct ends and often two different audiences: they provide crucial datasets for NLP projects such as automatic parsing and grammar induction while also providing a valuable corpus for scholars researching the state of a language and its progression across time.

Historical treebanks, however, also offer one additional benefit over modern treebanks: they provide an annotated set of texts that scholars actually care about. When linguists of modern languages base theories on corpus evidence, their analysis is generally directed toward the language at large; few, if any, pore over the *Wall Street Journal* examining its use of an arcane literary device. If the corpus is Vergil, however, we do.

The sheer volume of Latin texts available electronically[1] - not to mention the enormous mass still locked in print - is much larger than the small community of scholars and students who can read it. This alone justifies a treebank as a resource for those attempting to learn the language, but it also highlights the need for automatic methods of parsing and machine translation. To this end a Latin treebank will well serve the NLP community, which has a long history of applying such research to modern languages.[2] Classical scholars, however, largely operate on a fixed canon of texts. The value of a treebank for them is not so much in training

[1] The open-access Perseus Digital Library [10] alone hosts a 3.4-million-word corpus of Classical Latin texts, while the private Biblioteca Teubneriana Latina [5] has a digital collection of 10 million words.

[2] See Carroll [7] for a summary of automatic syntactic parsing and Charniak et al. [8], Quirk et al. [26] and Yamada and Knight [29] for recent approaches to syntax-based MT.

Hajič, J. and Nivre, J. (eds.): Proceedings of the TLT 2006, pp. 67–78.

language models to deal with new, unseen, text, but in rigorously analyzing the text we have.

This paper, therefore, has two purposes: first, it introduces early work that has been completed to date on the creation of a Latin treebank; second, it illustrates the potential uses of such a treebank not only for linguists and computer scientists (who are already familiar with the value of large corpora), but for traditional Classical scholars as well.

2 Building a Latin Treebank

Latin is a heavily inflected language with a high degree of variability in its word order. Indeed, Latin exhibits two different varieties of "free" word order. One is variable headedness, so that, e.g., verbs can appear at the end of a sentence (SOV), at the beginning (VSO/VOS), or in the middle (SVO). The other is non-projectivity, where constituents themselves are broken up with elements of other constituents, as in the dependency graph shown in figure 1, where an arc drawn from *iram* to *memorem* crosses that drawn from *Iunonis* to *saevae* (*ob* is also non-projective).[3]

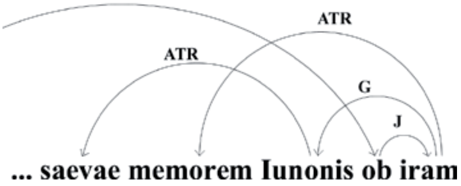

Figure 1: Dependency graph of *saevae memorem Iunonis ob iram*, "on account of the mindful anger of cruel Juno" (Vergil, *Aeneid* 1.4). Arcs are directed from heads to their dependents.

2.1 Annotation

This high level of non-projectivity has encouraged us to base our annotation style on that used by the Prague Dependency Treebank (PDT) (Hajič [13]) for Czech (another non-projective language) while tailoring it for Latin via the grammar of Pinkster [24]. Based on the dependency grammar of Mel'čuk [19], PDT annotation foregoes the non-terminal phrasal categories of context-free grammars, instead linking words themselves to their immediate heads. The full Latin tagset is given in table 1.

While our tagset is predominantly based on that used by the PDT, Latin presents a number of syntactic idiosyncrasies that necessitate a departure from it. These characteristics, along with a desire to tailor the treebank to enable certain research

[3] See Nivre [21] for a formal definition of projectivity.

PRED	predicate
SBJ	subject
OBJ	object
IOBJ	indirect object
COMP	complement
ATR	attribute
G	possessive genitive
REL	relative clause
ADV	adverbial
J	object of preposition
SubV	subordinate verb
X	non-coordinating punctuation
APOS	apposition
AUX	auxiliary verb
EXCLAM	exclamatory
PNOM	predicate nominal
SPCH	direct speech
ABS	ablative absolute
CO	coordination
ExD	ellipsis

Table 1: Complete Latin tagset.

questions, has caused us to annotate possessive genitives, a unique clausal construction called the "ablative absolute," and ellipsis in a manner slightly different than in the PDT.

2.1.1 Possessive Genitives

The PDT generally annotates possessive genitives and relative clauses with the "Atr" tag when such phrases attributively modify nouns. Our tagset splits this tag into three separate categories: ATR (for general attributive relationships, such as those of adjectives), G (for possessive genitives) and REL (for relative clauses). In Latin all adjectives match the case, number and gender of their head nouns; since most adjectives also have the ability to become substantival, if we assigned the same tag to possessive genitive relationships as to attributive relationships, we would have no way of distinguishing whether the phrase *liber pueri boni* means "the book of the good son" or the "the book of the son of the good man." A dedicated tag lets us do so.

2.1.2 The Ablative Absolute

The ablative absolute is a grammatical construction similar to the English nominative absolute, where a noun and (typically) a participle form a phrase that is disjoint from the grammar of the rest of the sentence; in Latin both the noun and participle are inflected in the ablative case, as in the following:

> His rebus cognitis Caesar Gallorum animos verbis confirmavit (Caesar, *De Bello Gallico* 1.33)
>
> These things understood, Caesar encouraged the minds of the Gauls with words.

Absolute constructions are not completely separated from the main clause, however, since the action of that clause is understood to take place in the context of the absolute (i.e., not coordinated with it). This would suggest the use of an adverbial tag, but using such a single tag would not allow us to distinguish between ablative absolutes (like *his rebus cognitis*) and other nouns, participles or adjectives that instrumentally modify a verb (as *verbis* modifies *confirmavit* above). A dedicated tag preserves the ability to isolate these constructions from the others.

2.1.3 Ellipsis

Ellipsis - the omission of words in a sentence that are recoverable from contextual cues - is also a pervasive phenomenon in these texts. Our method of representing ellipsis attempts to preserve the structure of the tree as much as possible. We accomplish this by assigning a complex tag to orphaned words. This tag preserves the path from the word itself to the elided word's head. Consider the example of *unam incolunt Belgae, aliam Aquitani* (Caesar, *De Bello Gallico* 1.1) given in figure 2.

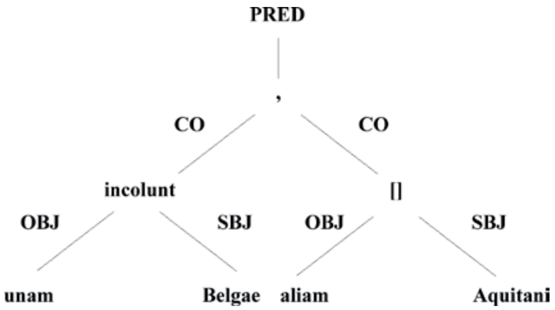

Figure 2: Dependency tree of *unam incolunt Belgae, aliam Aquitani* ("one the Belgians inhabit, another the Aquitani").

Here, the verb *incolunt* is missing from the second clause. We can preserve the structure of the tree by assigning the head of *aliam* and *Aquitani* to be the head that *incolunt* would have if it were in the sentence (the coordinating comma), and by assigning tags to each that preserve the path: *aliam* should be the object (OBJ) of *incolunt*, which should then depend on the coordinating comma via CO;

it therefore receives the tag OBJ_ExD0_CO (like the PDT, ExD here signifies an external dependency; the following numeral indexes the ellipsis, since in some sentences multiple words are elided). Likewise, *Aquitani* should be the subject (SBJ) of the elided word; it therefore receives the tag SBJ_ExD0_CO. This method allows us to use the complex tags to reconstruct the tree as necessary.

2.2 Treebank

Using this tagset, we have broken ground on a Latin Dependency Treebank by annotating excerpts from four texts, a total of 12,098 words, as distributed in table 2. The texts are Cicero's *Oratio in Catilinam* [2] (delivered ca. 63 BCE), Caesar's *Commentarii de Bello Gallico* [1] (published ca. 51 BCE), Vergil's *Aeneid* [4] (composed up to his death in 19 BCE) and Jerome's *Vulgate* [3] (composed ca. 405 CE). The size of this initial treebank is of course quite small, but it is intended as the seed of a much larger, million-word project.

Date	Author	Words
63 BCE	Cicero	1,189
51 BCE	Caesar	1,486
19 BCE	Vergil	2,647
405 CE	Jerome	6,776
	Total:	12,098

Table 2: Treebank composition by author.

In addition to the index of its syntactic head and the type of relation to it, each word is also annotated with the lemma from which it is inflected and its morphological code (a composite of nine different morphological features: part of speech, person, number, tense, mood, voice, gender, case and degree).

In order to bootstrap the treebank, we refined our annotation standards in the process of annotating this core sample of texts: at this early stage, the annotations were provided by a single annotator and validated by another; in the future, as we add more texts and expand the scope of the project, we plan to provide multiple annotations by independent annotators.

2.3 Non-Projectivity

While the following section describes the variety of linguistic experiments that can be performed on this annotated data, one initial observation is appropriate here: the non-projectivity rates of each author in our corpus. The non-projectivity rate of a language impacts, among other things, the accuracy of automatic parsing methods[4] - the higher the rate, the more difficult the task. Nivre and Nilsson [22] report a

[4] For two approaches to parsing non-projective languages, see Nivre et al. [23] and Collins et al. [9].

1.81% non-projectivity rate (by word) for Czech and .94% for Swedish using data from the PDT and the Danish Dependency Treebank. As table 3 shows, the rates we observe for Latin are much higher.

Author	Non-Projectivity rate (by word)
Jerome	1.8%
Caesar	2.9%
Cicero	5.8%
Vergil	12.2%

Table 3: Non-projectivity rates by author.

These extremely high rates are undoubtedly due to the highly literary nature of the texts and the importance of stylistic decisions on the surface word order. Jerome and Caesar are both (moderately) unembellished prose authors, Cicero a highly stylized orator, and Vergil a poet for whom the word order of a verse often complexly interacts with meter.

3 Using a Latin Treebank

A Latin treebank has the potential to be used as a knowledge source in a number of traditional lines of inquiry, including rhetoric, lexicography, philology and historical linguistics. To demonstrate the power of a corpus of syntactically annotated sentences, we will investigate two of these topics:

1. Many Latin rhetorical devices involve the interface between syntactic dependencies and word order. A treebank annotated with dependency relations would let us uncover and quantify these devices.

2. Classical Latin is thought to be predominantly characterized by SOV word order, but that order eventually transformed into the SVO typical of romance languages. A treebank comprised of texts spanning several centuries would let us measure the rate of this change.

Our preliminary work on creating a Latin Dependency Treebank provides an initial testbed on which to illustrate the value of a treebank for answering these kinds of questions. The sample size for these experiments is of course much too small for reliable statistics, but they nevertheless provide an interesting proof-of-concept for leveraging treebanks for Classical research.

3.1 Rhetorical devices

Since Latin is a highly inflected language, authors are free to exploit word order to achieve rhetorical effect. This is often taken to an obfuscating extreme in Classical poets, whose medium defines a more fluid and unpredictable space. As a rhetorical device, the transposition of order is known as hyperbaton, and classical rhetoricians such as Quintilian recognized both its utility and its potential for confusion: while noting that elegance often demands it, Quintilian[5] also finds fault in the following line of Vergil for being excessive:

> tris notus abreptas in saxa latentia torquet,
> saxa vocant Itali mediis quae in fluctibus aras. (Vergil, *Aen.* 1.108-9)
>
> three | the south wind | snatched | into | the rocks | lying | turns
> the rocks | call | the Italians | middle | which | in | waves | altars
>
> The south wind turns the three snatched (ships) into the rocks, such
> rocks (lying in the middle of the waves) the Italians call "altars."

Not all hyperbaton, however, is detrimental, and as witnessed by the projectivity rates given above, it is ubiquitous in poetry and also present in prose. Locating instances of hyperbaton in a treebank involves two tasks: asserting a projective order of words in a sentence, and discovering locations where the observed order departs from it. While a large-scale treebank will enable a reliable grammar to be induced from it, we can illustrate the process with the following example.

One of the more common forms of hyperbaton is the transposition of an adjective normally found within a prepositional phrase to a location outside of it (typically to the position immediately to the left of the preposition itself). This phenomenon occurs in such common phrases as *magna cum laude*, but also especially in poetry, as in figure 3.

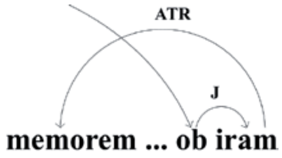

Figure 3: Dependency graph of *memorem* ("mindful") *ob* ("on account of") *iram* ("anger") ("on account of the mindful anger").

Canonical order would place the adjective *memorem*, which modifies the noun *iram*, inside the prepositional phrase headed by *ob*, as, for example, *ob iram memorem*.

A similar, but much less common, variety of prepositional hyperbaton involves transposing the object of the preposition itself outside of the prepositional phrase, as, for example, *iram ob memorem* in figure 4.

We can use a treebank to easily locate this kind of rhetorical device: for the former, we look for all words that occur to the left of a preposition but modify a word to the right of it that has the preposition as its head; for the latter, we look for all words whose immediate head is a preposition and that occur to the left of that head. Table 4 summarizes the results of this query.

[5] *Inst. Orat.* VIII.II.14.

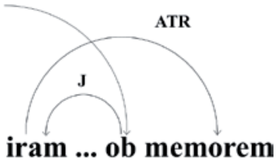

Figure 4: Dependency graph of *iram* ("anger") *ob* ("on account of") *memorem* ("mindful") ("on account of the mindful anger").

Author	*adj < prep < noun*	*noun < prep < adj*
Vergil	40.0%	15.6%
Cicero	8.9%	0%
Caesar	2.2%	0%
Jerome	0%	0%

Table 4: Prepositional phrase hyperbaton rates. *adj < prep < noun* hyperbaton is that of the form *memorem ob iram*; *noun < prep < adj* is that of the form *iram ob memorem*. Vergil, n=90; Cicero, n=45; Caesar, n=138; and Jerome, n=540.

This patterning falls in line with our intuitions: Vergil, the only poet of the group, exploits both varieties of hyperbaton far more frequently than the others; and of the prose authors, the more marked object transposition is non-existent, while adjective transposition occurs at a rate consistent with the author’s literary style: Cicero, a polished orator, frequently exploits rhetorical devices, while Caesar does not; Jerome, writing over four hundred years later and in a deliberately plain style, does not use such hyperbaton at all.

3.2 Word Order

Classical Latin word order is generally thought to be SOV,[6] but eventually transformed into the SVO order found in modern-day romance languages. Several studies have attempted to quantify the word order distribution of various authors for the purpose of comparison, such as Petronius (Hinojo [14]), Cicero (Pinkster [24]) and the Vulgate (Metzeltin [20]),[7] but are necessarily limited by the manual effort required to produce such data.

Treebanks, however, provide at least a first step in determining the predominant word order for a corpus (or any subset of it). In order to uncover the surface word order distributions of the four authors comprising our preliminary Latin treebank, we search the corpus for all verbs and locate the subject and object that depend on each. Note that any number of other, more restricted, searches are also possible, such as that for only simple finite verb forms, for only verbs within relative clauses,

[6] See, for example, Kühner and Stegmann [17] (§246) and Marouzeau [18], but see also Pinkster [25] for an argument that SOV evidence is lacking.

[7] See Pinkster [25] for a summary.

for only pronominal subjects, or even only for the verb *video*. The results in table 5 are for an unrestricted search of all verbs (both finite and non-finite) with no restriction on the type or part of speech of their dependents.

	Cicero	Caesar	Vergil	Jerome
SVO	5.3%	0%	20.8%	68.5%
SOV	26.3%	64.7%	18.8%	4.7%
VSO	5.3%	0%	6.3%	16.5%
VOS	0%	0%	10.4%	3.1%
OSV	52.6%	35.3%	25.0%	3.9%
OVS	10.5%	0%	18.8%	3.1%

Table 5: Word order ratios with full subjects and objects. Cicero, n=19; Caesar, n=17; Vergil, n=48; and Jerome, n=127.

Since repeated (or implied) subjects or objects can be realized as zero-anaphora in Latin, a high proportion of the sentences take the surface form of OV, VO, SV or VS. Tables 6 and 7 present these figures.

	Cicero	Caesar	Vergil	Jerome
OV	68.2%	95.2%	56.2%	13.9%
VO	31.8%	4.8%	43.8%	86.1%

Table 6: Word order ratios with a zero subject. Cicero, n=44; Caesar, n=63; Vergil, n=121; and Jerome, n=309.

	Cicero	Caesar	Vergil	Jerome
SV	75.9	86.7%	53.6%	65.8%
VS	24.1%	13.3%	46.4%	34.2%

Table 7: Word order ratios with a zero object. Cicero, n=58; Caesar, n=90; Vergil, n=97; and Jerome, n=404.

Caesar here presents good evidence for at least superficial SOV word order,[8] with 64.7% of his sentences coming in this form; the remaining 35.3% are instances of object fronting likely due to topicalization (the movement of a stressed phrase to the start of the sentence). Cicero displays more variety, but still suggests that the basic order is SV/OV, while Vergil offers a more diversified distribution. Jerome, writing over four hundred years later than the Classical authors, shows strong evidence for SVO order. The variety of these results reinforces the fact that statistical analysis is only a first step, and that further qualitative analysis is necessary. Treebanks are an indispensable tool for this task as well, allowing us to

[8] See Adams [6] for a typological argument (based on the universals of Greenberg [12]) that the frequency of OV in Caesar and Cicero is due not to a natural SOV order but rather the artificial imitation of one (as the hallmark of a higher register).

gather and compare any subset of the corpus (such as all OSV sentences in Caesar). In addition to providing quantitative evidence, treebanks are just as valuable for occasioning human-oriented analysis as well.

4 Conclusion

A corpus of syntactically parsed Latin sentences has the potential to impact both computational linguists and traditional Classical scholars. The nascent treebank presented here provides an interesting proof-of-concept for the variety of questions that can be answered with a large syntactic corpus - not only can it provide a quantified and reproducible answer to such questions, but it can also suggest and facilitate further qualitative analysis as well. This treebank is still in its infancy; in the future we plan to further annotate the canon of Classical texts and expand the variety of authors to include texts from the pre-Classical period (ca. 200 BCE) up to the Early Modern era (ca. 1780 CE). The treebank itself will be openly available to the public.

5 Acknowledgments

Grants from the Digital Library Initiative Phrase 2 (IIS-9817484) and from the National Science Foundation (BCS-0616521) provided support for this work.

References

[1] Caesar, C. Julius, *Commentarii Rerum in Gallia Gestarum VII; A. Hirti Commentarius VIII*. T. Rice Holmes (Oxford: Clarendon Press, 1914).

[2] Cicero, M. Tullius, *Orationes*. Recognovit brevique adnotatione critica instruxit Albertus Curtis Clark (Oxford: Clarendon Press, 1908).

[3] Jerome, *Vulgate Bible*. Bible Foundation and On-Line Book Initiative. ftp.std.com/obi/Religion/Vulgate.

[4] Vergil, *Bucolica, Aeneis, Georgica. The Greater Poems of Virgil*. J. B. Greenough (Boston: Ginn & Co., 1882).

[5] Biblioteca Teubneriana Latina, BTL-3 (Turnhout: Brepols; Munich: K. G. Saur, 2004).

[6] Adams, J. N. (1976), "A Typological Approach to Latin Word Order," *Indogermanische Forschungen* 81, pp. 70-99.

[7] Carroll, John (2003). "Parsing," in: Ruslan Mitkov (ed.), *The Oxford Handbook of Computational Linguistics* (Oxford: Oxford University Press), pp. 233-248.

[8] Charniak, Eugene, Kevin Knight, and Kenji Yamada (2003). "Syntax-Based Language Models for Machine Translation," *Proceedings of the MT Summit IX*.

[9] Collins, Michael, Jan Hajič, Eric Brill, Lance Ramshaw, and Christoph Tillmann (1999). "A Statistical Parser of Czech," *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (College Park, 1999), pp. 505-512.

[10] Crane, Gregory R., Robert F. Chavez, Anne Mahoney, Thomas L. Milbank, Jeffrey A. Rydberg-Cox, David A. Smith, and Clifford E. Wulfman (2001), "Drudgery and Deep Thought: Designing a Digital Library for the Humanities," *Communications of the Association for Computing Machinery* 44.5, pp. 35-40.

[11] Demske, Ulrike, Nicola Frank, Stefanie Laufer and Hendrik Stiemer (2004), "Syntactic Interpretation of an Early New High German Corpus," *Proceedings of the Third Workshop on Treebanks and Linguistic Theories (TLT 2004)*, pp. 175-182.

[12] Greenberg, Joseph H. (1963), "Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements." In: Joseph H. Greenberg, *Universals of Grammar* (Cambridge: MIT Press), pp. 73-113.

[13] Hajič, Jan (1999), "Building a Syntactically Annotated Corpus: The Prague Dependency Treebank," in: E. Hajičová (ed.), *Issues of Valency and Meaning. Studies in Honour of Jarmila Panevová* (Prague: Charles University Press), pp. 106-132.

[14] Hinojo, G. (1985), "Del Orden de Palabras en al Satiricon," in: J. L. Melena (ed.), *Symbolae Ludovico Mitxelena Septuagenario Oblatae* (Vitoria: Vitoria University Press), pp. 245-254.

[15] Kroch, A., and A. Taylor (2000), Penn-Helsinki Parsed Corpus of Middle English, second edition. http://www.ling.upenn.edu/hist-corpora/PPCME2-RELEASE-2/

[16] Kroch, A., B. Santorini, and L. Delfs (2004), Penn-Helsinki Parsed Corpus of Early Modern English. http://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-1

[17] Kühner, Raphael, and Carl Stegmann (1914), *Ausführliche Grammatik der Lateinischen Sprache* II: *Satzlehre* (Hannover: Verlag Hahnsche Buchhandlung).

[18] Marouzeau, J. (1949), *L'Ordre des mots dans la phrase latine* (Paris: Les Belles Lettres).

[19] Mel'čuk, Igor A., *Dependency Syntax: Theory and Practice* (Albany: State University of New York Press, 1988).

[20] Metzeltin, Michael (1987), "Lateinische versus Romanische Satzgliederung?" in: Wolfgang Dahmen, Günter Holtus, Johannes Kramer and Michael Metzeltin (eds.), *Latein und Romanisch*, ser. Romanistisches Kolloquium I (Tübingen: Narr), pp. 246-69.

[21] Nivre, Joakim (2006), "Constraints on Non-Projective Dependency Parsing," *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 73-80.

[22] Nivre, Joakim, and Jens Nilsson (2005), "Pseudo-Projective Dependency Parsing," *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pp. 99-106.

[23] Nivre, Joakim, Johan Hall, Jens Nilsson, Gülsen Eryigit and Svetoslav Marinov (2006). "Labeled Pseudo-Projective Dependency Parsing with Support Vector Machines," *Proceedings of the Tenth CoNLL*, pp. 221-225.

[24] Pinkster, Harm (1990), *Latin Syntax and Semantics* (London: Routledge).

[25] Pinkster, Harm (1991), "Evidence for SVO in Latin?" in: Roger Wright (ed.), *Latin and the Romance Languages in the Early Middle Ages* (University Park, The Pennsylvania State University Press), pp. 69-82.

[26] Quirk, C., A. Menezes and C. Cherry (2005). "Dependency Treelet Translation: Syntactically Informed Phrasal SMT," *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pp. 271-279.

[27] Rocio, Vitor, Mário Amado Alves, J. Gabriel Lopes, Maria Francisca Xavier and Graça Vicente (2000), "Automated Creation of a Medieval Portuguese Partial Treebank," in Anne Abeillé (ed.), *Treebanks: Building and Using Parsed Corpora* (Dordrecht: Kluwer Academic Publishers), pp. 211-227.

[28] Taylor, Ann, Anthony Warner, Susan Pintzuk and Frank Beths (2003). York-Toronto-Helsinki Parsed Corpus of Old English Prose. University of York.

[29] Yamada, Kenji and Kevin Knight (2001). "A Syntax-Based Statistical Translation Model," *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pp. 523-530.