Kristin Bech and Kristine Gunn Eide

The annotation of morphology, syntax and information structure in a multilayered diachronic corpus

This paper describes the annotation scheme we use for old Germanic and Romance languages, with particular focus on syntax and information structure, and the issues of economy and disambiguation. We also discuss some of the annotation problems we have had to solve, in order to demonstrate the complexity of this kind of linguistic annotation.

1 Introduction

This paper reports on annotation work in the project Information Structure and Word Order Change in Germanic and Romance Languages (ISWOC[1]). In the project, we annotate morphology, syntax and information structure in a corpus consisting of texts from old Germanic (English, Norse/Norwegian, German[2]) and Romance (French, Spanish, Portuguese) languages. The purpose of the project is to study the relation between word order change and information structure in these languages, as all these languages had verb-second structures at some stage, but have developed in different directions. Multilayered means that the application is designed in such a way that the different levels build on each other, i.e. the application makes guesses for syntax on the basis of the morphological annotation, and the syntactic annotation in turn provides suggestions for which elements should be included in the information structure annotation (cf. HAUG ET AL. 2009). The aim of this paper is to describe the annotation principles, both from a theoretical and practical point of view, and discuss some of the problems we encounter in the annotation. We focus particularly on syntactic annotation and information structure annotation, and on how to balance the sometimes conflicting requirements of economy and disambiguation.

2 Economy vs disambiguation

Ideally, an annotation would enable us to retrieve *all* the examples of the particular phenomenon we are looking for as well as *only* the examples we are looking for, with as uncomplicated a search string as possible. However, considering the formal and structural ambiguities that we find in all languages, this ideal is not easily attainable. Also, since the annotation is a tool to enable linguistic research rather than a linguistic analysis in itself, it is not the ultimate goal, nor is it feasible, to disambiguate every ambiguous expression. When we evaluate whether or not to disambiguate, we must ask ourselves (1) to what extent that expression or construction will be retrievable through other means, (2) how much it will cost in terms of working time (disambiguation is very time-consuming) and (3) whether or not the ambiguity is a reflection of accidental formal likeness (for instance the Portuguese homonyms (a) reflexive pronoun *se* and (b) subjunction *se* 'if'), or whether it reflects a "constructed" ambiguity, perhaps because the grammatical categories that we use are in themselves less clear-cut than we would like them to be (cf. the discussion concerning demonstratives in section 3). Since we are constructing a multilayered corpus, we have to ask ourselves if, for instance, a certain formal morphological ambiguity will be disambiguated in the syntax layer, or the other way around, if syntactic ambiguity can be resolved in the

morphology layer, since there is little gain in meticulously disambiguating something in one layer that can easily be retrieved in another. We also have to consider whether it is at all possible to disambiguate; in many cases it is not possible.

3 Morphological annotation

The first level of annotation is morphology. In some cases we manage to import morphological annotation from other electronic corpora, but the annotation still involves a considerable amount of manual work. The challenges we experienced with regard to morphology typically had to do with what word classes to include, and with the fact that some words are ambiguous, or not easily classifiable, in terms of word class. In addition, since we wish to compare Germanic and Romance languages, it was important to keep the morphological specifications as consistent as possible. However, in some cases, we had to differentiate. One example is the distinction between demonstrative pronouns and demonstrative determiners. In Old Norse and Old English, there is no morphological distinction between these two categories, and demonstratives are always tagged as determiners, whether they are used as attributes or pronouns. Hence, there is no morphological disambiguation between the pronominal use of *þá* in (1) and the determiner in (2), both examples from Old Norse.

(1) ec vil tala við *þá* (*Strengleikar*, 13th c.)
I will speak with *them*.ACC.PL.M/*her*.ACC.SG.F

(2) um *þá* daga var þar iafnan ufriðr ok bardagar (*Strengleikar*, 13th c.)
in *those* days was there often unrest and war
'In those days there was often unrest and war'

The difference between *þá* in (1) and *þá* in (2) will rather appear in the syntactic annotation, where *þá* in (1) belongs to an argument category and *þá* in (2) is analyzed as an attribute.

Portuguese and Spanish have the same formal identity between demonstrative pronouns and determiners, but in these languages there is also a class of non-inflectional demonstrative pronouns that never modify nouns: *isto/esto* 'this' (1st person), *isso/eso* 'that' (2nd person), *aquilo/aquello* 'that' (3rd person). They are usually classified as neuter pronouns, as opposed to their inflected counterparts, which are used either pronominally or as determiners. We have chosen to keep these non-inflected pronouns together with the inflected pronouns/determiners in the same combined pronoun/determiner class that we use for Old English and Old Norse. In our view, the advantage of maintaining a common analysis for all the languages and the advantage of keeping the demonstratives in one group outweigh the problem of a tagging of three pronouns that is strictly speaking not correct. The non-inflected pronouns are retrievable because they are morphologically marked as uninflected.

4 Syntactic annotation

4.1 Some aspects of the syntactic model

Our syntactic model is based on dependency grammar (DG), as laid out by TESNIÈRE (1959).[3] The main idea of DG is that syntactic structure consists of lexical elements which are linked by asymmetrical relations called dependencies (see NIVRE 2005, chapter 3, for a nice overview of DG). Thus, DG, unlike phrase structure grammar, does not operate with phrasal nodes, as it is the relation between the head and its dependent(s) that determines

what the structure is. A dependent does not have to be adjacent to its head; dependency relations have to do with structural order rather than linear order. In our annotation application, word order is not modelled, but word order is retrievable since each token is indexed with a number showing its position in the sentence. Thus, we can search for features including word order, e.g. "all initial adverbial prepositional phrases followed by the verb and the subject". In the annotation of older languages (and certainly some modern languages as well), it is an advantage not to be constrained by word order, since word order was freer, i.e. used to mark information structure relations rather than grammatical relations, and discontinuous dependency frequently occurred. The tree diagram below (Figure 1) represents the syntactic annotation of the Old English sentence in (3).

(3) ac he ne mæg for scame in gan buton scrude (*Apollonius of Tyre*, 11th c.)
but he not may for shame in go without clothing
'but for shame, he may not enter without clothing'

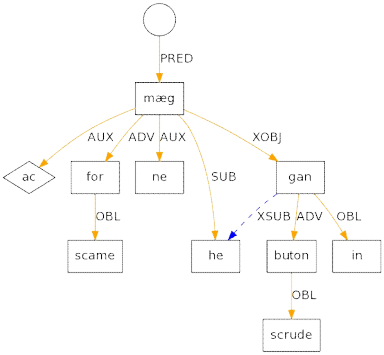

Figure 1: the dependency tree of example (3).

Three features of the tree structure should be noted. The first is the lack of empty nodes; each node corresponds to a word. It happens, however, that we have to insert an empty node, i.e., we may have to insert an empty conjunction in cases of asyndetic conjunction, or an empty verbal node, and the application is designed for and allows such cases. The second feature is the XOBJ relation, which may be unfamiliar from a traditional grammar point of view. This relation is used for predications that have external subjects (they get the subject via coreference relations within the sentence), and are governed by another verb. Non-finite verbs are typically involved, but nouns and adjectives may also be in an XOBJ relation to the matrix verb, in predicative clauses. The XOBJ elements have to be selected and demanded by the matrix verb and are thus arguments. There is also another relation (not shown in Figure 1) which involves adverbials with an external subject: XADV. This relation is typically found with present participles, as in the Old English sentence in (4), where *bysmriende* 'mocking' is XADV to *cwædon* 'said' and has the external subject *heahsacerdas* 'chief priests', which it shares with the matrix verb.

(4) Eallswa þa heahsacerdas bysmriende betwux þam bocerum cwædon
(*Old English Gospels*, 11th c.)
likewise the chiefpriests mocking between the scribes said
'Likewise also the chief priests mocking said among themselves with the scribes'
(*Authorized King James Version*)

This brings us to the third feature: The syntactic model we have adopted differs markedly from DG in one particular respect, namely in allowing structure-sharing, similar to Lexical-Functional Grammar. In DG, a word can only have one head, but this principle leads to problems in the treatment of non-finite verbs in particular. In (3), *he* 'he' is the subject of both *mæg* 'may' and *gan* 'go'. In our graph, this is represented with the slashed arrow, which shows that *he* is the external subject, XSUB, to *gan*. We also use these secondary edges to show other types of shared arguments, most notably in the very frequent case of an ellipted subject in coordinate constructions: *He came, saw and conquered*, where *saw* and *conquered* would both slash to the subject *he*.

4.2 Some syntax annotation challenges

4.2.1 Complex verb phrases

One conflict between the languages in our corpus relates to the status of auxiliary[4] verbs. While the old Germanic languages have a rather limited set of auxiliaries, the Romance languages have a much larger group of verbs that may function as auxiliaries in verbal periphrases. For all the languages, we analyze auxiliary verbs that are used to mark tense as an AUX[5] element attached to the head (and not the other way around) and attach the arguments directly onto the non-finite verb, as in the Old Norse example in (5). Furthermore, Old Norse combines modality and tense, using modal verbs to form futures and conditionals. Consequently, temporal and modal verbs are therefore not distinguished, since disambiguation is not possible. Old Norse therefore has the same annotation for the tense-marking verb *hafa* 'to have' (5) and modal verbs, e.g. *vilia* 'to wish to' in (6).

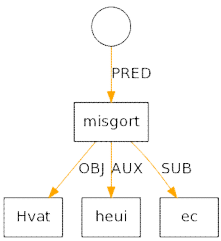

Figure 2: the dependency tree of example (5).

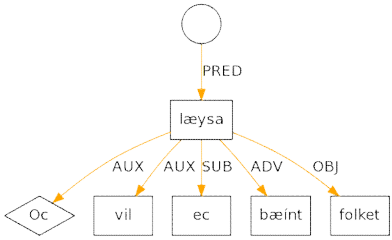

Figure 3: the dependency tree of example (6).

(5) Hvat heui ec misgort (*Strengleikar*, 13th c.)
what have I misdone
'What have I done wrong?'

(6) Oc vil eg beínt læysa folket (*Óláfs saga hins helga*, 13th c.)
and will I straightaway release people.DEF.ART
'And I will straightaway release the people'

In Old English, on the other hand, examples equivalent to the Old Norse example in (6) receive a different analysis, i.e. an analysis with XOBJ (see example (3)), since it can be argued that the verbs that today are considered as modal auxiliaries still had at least some of their lexical force in Old English (TRAUGOTT 1992). As regards the syntactic annotation, then, the Old Norse structure is flat, whereas the analysis of Old English is biclausal.

We analyze the Romance tense and modal auxiliaries in the same way as their Old English counterparts, but the Romance languages also have a range of more or less grammaticalized aspectual auxiliaries that combine with the main verb, and are often preceded by a preposition. These aspectual auxiliaries have given us some headache. There is variation as to what extent these aspectual auxiliaries have been grammaticalized, from the almost completely grammaticalized verb (in Portuguese) *estar* in *estar fazendo* or *estar a fazer* 'to be doing something', through semi-grammaticalized verbs such as *passar* in *passar a fazer* 'to pass on to doing something (else)' to non-grammaticalized finite verbs like *começar* in *começar por fazer* 'start by doing'. In the case of *estar*, which is similar to its English counterpart *be* in progressive constructions, it is relatively easy to argue that the verb is no more than an auxiliary in periphrastic constructions; it has no semantic content other than tense and aspect. With the other verbs, however, we are dealing with a grammaticalization continuum, so if we were to analyze the periphrases according to their grammaticalization status, with the finite verb either as an AUX or as a head, the choice would in most cases be rather arbitrary. We therefore chose to use the same analysis for all these periphrases, regardless of grammaticalization status. This means that we annotate both the subject and the non-finite verb as arguments to the finite verb, and indicate the subject of the infinitive through a slash annotation as described in 4.1.

Our analysis of modals has has consequences for retrievability later on. In Old Norse, modals are analyzed in the same way as other auxiliaries, whereas in our other languages, modals are analyzed as heads in an XOBJ relation, as we do with other verbal periphrases. Hence, the modals will only be retrievable through a search on language-specific lemmas or through relatively complicated syntactic and morphological search strings. A search asking for a particular syntactic structure, e.g. "find a finite verb with an XOBJ that is an infinitive" would render not only the modal verbs, but also other periphrases with the same structure, e.g. the Romance aspectual verbal periphrases described above, or the Old English causative constructions with an accusative + infinitive described in 4.2.2. Also, whether we search for lemmas or structures, the search strings will have to be language-specific. In other words, though some serious thinking will be required in the search stage, the annotation stage is kept simple and manageable for the annotators.

4.2.2 Extended Case Marking and other infinitival clauses

In section 4.1, we described the XOBJ relation where the finite and the non-finite verb share a subject, and this is tagged through slash annotation. The subject of the non-finite verb does not have to be the same as that of the finite verb, but it must be external for a relation to be analyzed as XOBJ; if it is internal we analyze the sentence as a complementizer clause through the relation COMP.

Accusative with infinitive (AcI) constructions are typically COMP, since the accusative + infinitive unit is regarded as an argument of the finite verb, and the subject gets its semantic (theta-) role from the non-finite verb. Example 7 is a case in point, with the syntactic representation shown in Figure 4.

(7) oc hævir hann hœyrt [fugla syngia]... (*Strengleikar*, 11th c.)
and has he heard birds.acc sing.inf...
'and he has heard birds sing...'

(8) oc bauð [þeim] [at biða sin] (*Strengleikar*, 13th c.)
and bid them.dat to wait her.refl
'and bid them wait for her'

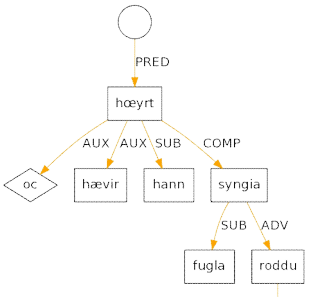

Figure 4: the dependency tree of the accusative with infinitive in example (7).

In Old Norse, there is a distinction between accusative with infinitive after verbs of perception (example 7), and dative with infinitive after verbs of causing and commanding such as 'to order', 'to bid', 'to let', 'to make' (someone (to) do something) (example 8). The latter is analyzed as an XOBJ construction in which the dative is an argument of the finite verb and is tagged as an external subject to the non-finite verb through a slash annotation.

In Old English, both perception verbs and verbs of causing and commanding are followed by accusative + infinitive. It is therefore difficult to distinguish between the constructions; i.e. which ones are true AcI constructions in which the AcI as a unit is dependent on the finite verb, and which ones are constructions in which the accusative and the infinitive are separate arguments of the finite verb. MITCHELL (1985 II) discusses AcI extensively, and makes repeated use of the 'jungle' metaphor to describe the difficulty of the task. In any case, we have decided to analyze Old English in the same way as Old Norse, with perception verbs taking AcI, and thus COMP, and causative verbs and verbs of commanding taking an XOBJ construction; in other words, the accusative element is in the latter case analyzed as an argument of the finite verb. The motivation for this is the existence of various constructions for a verb such as *hatan* 'to order, command, bid', such as the ones exemplified in (9) and (10). Especially the latter construction suggests that the accusative element is in fact an argument of the finite verb.

(9) and het [hine] [in gan] (*Apollonius of Tyre*, 11th c.)
and bade him.ACC in go
'and bade him enter'

(10) þa het ic [eallne þone here] [þæt he to swæsendum sæte]
(*Letter of Alexander the Great to Aristotle*, 10th-11th c.,
from MITCHELL 1985 vol. II: 871)
then bid I all.ACC the.ACC army.ACC that it to meal sit.SBJV
'then I bid the entire army to be seated for the meal'

As regards the Romance languages, the distinction between the COMP and XOBJ construction is not easily made, because the dative/accusative distinction is less clear. We have therefore chosen to analyze both constructions with perception verbs and constructions with causative verbs as COMP constructions. At first sight, a consistent COMP analysis may seem like the poorer alternative, because the syntactic relation between the finite verb and the subject of the infinitive is lost with this annotation. There are, however, other reasons for opting for the COMP analysis.

While in most of our languages, there is a distinction between accusative and dative subjects of the infinitive, these constructions appear in three different structures in Portuguese (MARTINS 2006): non-inflected infinitive with accusative subject (Extended Case-Marking (ECM) construction), as in (11), inflected infinitive with nominative subject, as in (12), and non-inflected infinitive with dative subject, as in (13). All three sentences mean 'I ordered the students to make the cake'.[6]

(11) Mandei os alunos fazer o bolo
I ordered the students make.INF the cake

(12) Mandei os alunos fazerem o bolo
I ordered the students make.INF.3PL the cake

(13) Mandei aos alunos fazer o bolo
I ordered to-the students make.INF the cake

The example in (12) does not have the structural ambiguity of the two others, because there is agreement between the (nominative) subject of the infinitive and the infinitive itself. Consequently, (12) must be analyzed as [mandei [os alunos fazerem o bolo]], with the subject case-marked by the infinitive and not by the finite verb (MARTINS 2006). For (11) and in particular for (13) there are two possible analyses: [mandei [(a)os alunos] [fazer o bolo]] and [mandei [(a)os alunos fazer o bolo]]. The easiest way to capture this structural ambiguity, would be to use an XOBJ relation for the infinitive, annotate the subject of the infinitive as an argument to the finite verb and use a slash to indicate the subject relation. Examples (11) and (13) would then be distinguished from (12), where the infinitive clause is analyzed as COMP. The main reason why we have chosen not to do so is that the third person *singular* form of the inflected infinitive is formally identical to the non-inflected form. In the singular, both (11) and (12) would be *mandei o aluno fazer o bolo* 'I asked the student to make the cake'. Because the third person singular is the most common form, there will be a large

number of ambiguous contexts where it is impossible to distinguish between a nominative + inflected infinitive and an accusative with infinitive.

A disadvantage of the consistent COMP analysis for Portuguese is the irretrievable status of the accusative or dative NP as object/indirect object to the finite verb, once the NP is analyzed as a subject of the infinitive. One consequence of this is that when we carry out the information structure analysis (as described in section 5), the information structure annotation on these arguments will be retrievable as information structure on subjects, not on objects. For example, if we search for "main clause objects that constitute given information", none of the AcI NPs will be included in the result. On the other hand, this can also be an advantage, since we will get fewer examples that may have to be removed anyway because of their ambiguous nature. Another advantage of the COMP analysis is that these constructions are the only COMPs with an infinitive as head. A simple search asking for a "COMP with a head that is infinitive" will render all and only these structures. Given their particular ambiguous nature, in Portuguese at least, it is a good idea to keep them apart from other structures. In addition, there is the additional advantage of keeping the annotation work simple and economical.

The fact that the subcorpora of different languages are annotated in basically the same way facilitates searches to a certain extent. However, in sections 3 and 4 we have seen that one and the same phenomenon may receive different analyses in different languages due to language-internal classification problems, and thus language-specific search strings are required. The information structure annotation, on the other hand, is not language-specific, as the categories are not tied to language-specific properties such as e.g. word order or definiteness. Rather, the scheme for information structure annotation is a method of textual analysis that may be applied to texts in any language, without the language-specific rules needed for morphology and syntax, as will be shown in section 5.

5 Information structure annotation

5.1 The information structure annotation scheme

The information structure annotation aims to represent the writer's assumptions about what is in the mind of the addressee (cf. e.g. CHAFE 1976, PRINCE 1981), and we are interested in how these assumptions contribute to linguistic structure, in particular word order. We use the PROIEL annotation scheme (cf. PROIEL guidelines,[7] and also NISSIM ET AL. 2004, and RIESTER ET AL. 2010) and annotate noun phrases only, distinguishing between old, new and accessible information on one level, and specific, non-specific and generic information on another.

An element is tagged as OLD information if it is co-referential with an element in the preceding discourse. If the old element is outside the limit of the annotation window (13 sentences), it is tagged as OLD-INACTIVE. An element is tagged as NEW if it is mentioned for the first time. The category Accessible subsumes three types of accessible information, namely situational (ACC-SIT), inferable (ACC-INF) and general (ACC-GEN) information. ACC-SIT is used for referents that are available in the discourse situation (mostly relevant in direct speech contexts), ACC-INF is used when the referent is inferable from the preceding discourse, and ACC-GEN refers to world knowledge, i.e. what we think was the world knowledge of the readers of the texts at the time when they were produced. OLD and ACC-INF elements must be textually licensed; they receive an anaphoric link which points back to the licensing element.

The OLD, ACC and NEW categories were the original ones in the annotation scheme, and they all have specific reference. However, the need arose to tag specificity vs. non-specificity as well, and thus the economy–disambiguation scale had to be tilted somewhat in favour of disambiguation. Annotators now have more categories to consider, but the advantage is a more satisfactory and, it is hoped, useful end product. Non-specific referents are basically divided into quantifier restrictions, QUANT, and all other types of non-specific referents. Furthermore, non-specific referents can be old or inferred. In the sentence *All people have nails and have to cut them*, *people* is quantified and receives the QUANT tag, *nails* is NON-SPEC and *them* is NON-SPEC-OLD (cf. PROIEL guidelines for a discussion of these categories and some other special cases). There is also a tag, KIND, for generic expressions, such as *lions* in *Lions are dangerous*.

Before we look at some examples of actual annotation, it should be mentioned that we also annotate NPs which constitute the complement of a preposition, e.g. *that house* in the prepositional phrase *in that house*. In addition, we annotate pro-dropped arguments, which are most commonly subjects, but pro-dropped objects and obliques also occur. Thus, in the information structure layer of the annotation application, we insert a marker for pro. Elements can refer back to pro, and pros can thus be elements in an anaphoric chain.

In the information structure annotation, we assume that certain textual properties, such as givenness, and certain semantic properties, such as specificity, are related to information structure properties such as topic, background and focus, and that the annotation will enable us to detect the information structure system in a given language.

5.2 Some information structure annotation challenges

In this section, we focus on two annotational issues that tend to cause a certain degree of frustration for the annotators: the Accessible category, in particular how to recognize inferables (ACC-INF), and complex noun phrases, i.e. noun phrases that contain more than one NP. One of the main questions as regards inferables is what we can assume was inferable for the intended readers of the text; it may very well be that contemporary readers of the old text inferred different things from the discourse than we do today. What makes this task a little less difficult is that the text guides us; the inferred elements have to be textually licensed from some element in the text, though not necessarily a co-referential element. As a warm-up exercise, let us first consider the entire information structure annotation of (14), which is the first verse of chapter 15 in the Old English Gospel of Mark. To provide some context, the last verse of the previous chapter is given (in Modern English (*King James*) for the sake of simplicity). The annotatable elements are marked in grey.

(14) [And the second time the cock crew (OE: *þa eft sona creow se hana*). And Peter called to mind the word that Jesus said unto him, Before the cock crow twice, thou shalt deny me thrice. And when he thought thereon, he wept.]
þa sona on mergen worhton þa heahsacerdas hyra gemot mid ealdrum and bocerum and eallum werodum and PRO-SUB læddon þone hælend gebundenne and PRO-SUB sealdon him Pilato. (*Old English Gospels*, 11th c.)

then immediately in morning held the chiefpriests their council with elders and scribes and all council and led the saviour bound and delivered him Pilate.DAT
'then straightway in the morning the chief priests held a consultation with the elders and scribes and the whole council, and bound Jesus and delivered him to Pilate' (*King James Authorized Version*)

First, the prepositional complement *mergen* is tagged as ACC-INF. This is because in the last verse of the previous chapter, it is mentioned that the cock crowed: *þa eft sona creow se hana*. Consequently, the morning can be inferred, since the crowing of the cock signals morning; hence the anaphoric link goes from *mergen* back to the verb *creow*. This is a relatively straightforward example of an inferable relation. *Heahsacerdas*, *ealdrum*, *bocerum*, and *werodum* are all tagged as OLD-INACTIVE. They were mentioned in the previous chapter, so the information is old, but outside the annotation window of 13 sentences. *Hyra gemot* is a complex NP consisting of a possessive pronoun and a noun. The pronoun is tagged as OLD, with an anaphoric link back to *heahsacerdas* and *gemot* is tagged as NEW. Then we have a new main clause without an expressed subject, and in the information structure annotation, we insert a PRO-SUB before the verb *læddon*. The question is what tag this pro element should get. Is it old, i.e. coreferent with *heahsacerdas*, meaning that the chief priests led Jesus, or should it rather be interpreted as inferable, ACC-INF, from the chief priests, since it is probable that the priests had some minions to do the binding and leading for them? Here we choose the former solution, since we do not want to overinterpret the text. The next taggable element is *hælend*, which is OLD information here, because it links back to *me* in the previous chapter. Then we get another PRO-SUB, which is tagged as OLD, linking back to the previous PRO-SUB. *Him* is also OLD, with a link to *hælend*, and finally *Pilato* is tagged as ACC-GEN, since we assume that Pilate was generally known to the readers of the Bible.
With the exception of the pro element mentioned, the annotation of this sequence is straightforward. In (15), however, things get a little more complicated.

(15) and hi tosomne eall werod clypedon ... and beoton hine on þæt heofod mid hreode and spætton him on and heora cneow bigdon. (*Old English Gospels*, 11th c.)
and they together all band called ... and beat him on the head with stick and spit him on and their knees bent
'and they call together the whole band ... and beat him on the head with a stick and spit on him and bent their knees' (*King James Authorized Version*)

There are several annotatable elements here, but we will focus on *þæt heofod* and *heora cneow*, both NPs consisting of a head and a dependent. In *þæt heofod* we annotate *heofod* as ACC-INF, since the inference here is from humans to body parts; we can infer that people have heads. The demonstrative pronoun *þæt*, used in an article-like way here, is not referential, and thus not annotated. *Heora cneow* also refers to a body part, but this NP consists of the possessive determiner *heora*, which is referential, and the noun *cneow*. Here, the referent can be identified by means of an element within the noun phrase, namely *heora*, which gets the OLD tag, and *cneow* is therefore annotated as NEW, even though it, like *heofod* is inferable. The point is to distinguish between NPs which contain the element we need to identify the referent, and NPs that do not contain any such element. In *heora cneow*, we get that information through *heora*, which refers back to *hi* 'they', and hence *cneow* is tagged as

NEW, whereas in *þæt heofod*, we do not get that information, and hence the head gets the ACC-INF tag.

Another example is found in (16). Again there are several annotatable elements, but we will focus on two of them.

(16) Þa wæs underntid and hie ahengon hine. And ofergewrit his gyltes wæs awriten, Iudea cyning (*Old English Gospels*, 11th c.)
Then was third-hour and they crucified him. And superscription his.GEN accusation.GEN was written, Jews.GEN king
'And it was the third hour, and they crucified him. And the superscription of his accusation was written over, THE KING OF THE JEWS'
(*King James Authorized Version*)

Here, *ofergewrit* 'superscription' is tagged as ACC-INF, with a link back to *ahengon* 'crucified', the inference being that superscriptions saying what the crime was were sometimes attached to the cross, and the readers would know this. *Gyltes* is also inferable, but this NP also contains a possessive determiner *his*, which identifies the referent; the link is back to the pronoun *hine*. *Gyltes* is therefore tagged as NEW; in other words, this NP is tagged in the same way as *heora cneow* in (16).

This type of annotation causes some problems for the annotators, because it is easy to overlook such inferences, it may be difficult to know which inferences are valid, and the tagging with old and new within the same NP (as in *his gyltes*) is not particularly intuitive. In other words, here economy has been sacrificed on the altar of disambiguation.

6 Summary

In this paper we have presented some aspects of our work on a corpus consisting of several languages in their older stages. We have given a brief description of the annotation scheme, focusing on syntax and information structure, and discussed some of the problems we come across in the actual annotation, with particular reference to the challenge of balancing out the principles of economy and disambiguation.

In the end, when we get to the search and research stage, we will be able to combine the annotated data of the three levels of morphology, syntax and information structure, and trace not only what types of arguments appear in which linear positions, but also the distribution of information structural properties within and between sentences, the research aim being to study the relation between information structure and word order change.

[1] *http://www.hf.uio.no/ilos/english/research/projects/iswoc/index.html*.

[2] We do not annotate German ourselves, since much work has been done on Old High German by our partners at the Humboldt University, Berlin, in the project *Informationsstruktur und Wortstellung im Germanischen*: *http://www2.hu-berlin.de/sprachgeschichte/forschung/informationsstruktur/index.php*.

[3] The model and application we use were developed by the PROIEL project at the University of Oslo, and we have adapted it for our languages. The PROIEL annotation guidelines can be found at: *http://folk.uio.no/daghaug/syntactic_guidelines.pdf*.

[4] We use the term 'auxiliary verb' here simply to refer to more or less grammaticalized verbs that modify the main verb in complex verb phrases. There is a gradience of auxiliary-hood among the verbs discussed here.

[5] AUX is the tag we use for grammatical words, i.e words that give additional information about their head. The tag is not only used for auxiliary verbs, but also marks modal particles, focus particles, negation, and coordinating conjunctions.

[6] The case of the infinitive subject is perhaps easier to see if we substitute *(a)os alunos* with pronouns: (11') mandei-os.ACC fazer o bolo; (12') mandei eles.NOM fazerem o bolo; (13') mandei-lhes.DAT fazer o bolo.

[7] *http://folk.uio.no/daghaug/info_guidelines.pdf*.

References

CHAFE, W. (1976). "Givenness, contrastiveness, definiteness, subjects, topics, and point of view." In LI, C. (ed.). Subject and Topic. New York: Academic Press, 25-55.

NIVRE, J. (2005). Inductive Dependency Parsing. Dordrecht: Springer.

HAUG, D.T.T, JØHNDAL, M.L., ECKHOFF, H.M., WELO, E, HERTZENBERG, M.J.B., MÜTH, A. (2009). "Computational and linguistic issues in designing a syntactically annotated parallel corpus of Indo-European languages." In Traitement Automatique des Langues, vol. 50, 17-45. Retrievable at *http://www.atala.org/Computational-and-Linguistic*.

MARTINS, A.M. (2006). "Aspects of infinitival constructions in the history of Portuguese." In GESS, R.S. & D. ARTEAGA (eds.). Historical Romance Linguistics: Retrospective and Perspectives. Amsterdam & Philadelphia: John Benjamins, 327-355.

MITCHELL, B. (1985). Old English Syntax, vol. II. Oxford: Clarendon Press.

NISSIM, M., DINGARE, S., CARLETTA, J, STEEDMAN, M. (2004). "An annotation scheme for information status in dialogue." In LINO M.T ET AL. (eds.) Proceedings of the 4th Interna tional Conference on Language Resources and Evaluation, Lisbon 2004. Retrievable at *http://homepages.inf.ed.ac.uk/jeanc/nissim-lrec2004.pdf*.

PRINCE, E. (1981). "Toward a taxonomy of given-new information." In COLE, P. (ed.). Radical Pragmatics. New York: Academic Press, 223-255.

RIESTER, A., LORENZ, D, SEEMANN, N. (2010). "A recursive annotation scheme for referential information status." In CALZOLARI, N. ET AL. (eds.). Proceedings of the 7th International Conference on Language Resources and Evaluation, Valletta 2010, 717-722. Retrievable at: *http://www.lrec-conf.org/proceedings/lrec2010/index.html*.

TESNIÈRE, L. (1959). Eléments de syntaxe structurale. Editions Klincksieck.

TRAUGOTT, E. C. (1992). "Syntax." In HOGG, R. (ed.) The Cambridge History of the English Language, volume I. Cambridge: Cambridge University Press, 168-289.