**Title:** The problem of syntactic ambivalence in corpus linguistics

**Author:** Ireneusz Kida

**Citation style:** Kida Ireneusz. (2012). The problem of syntactic ambivalence in corpus linguistics. "Lingua Posnaniensis" (Vol. 54, iss. 1, (2012), s. 57-63), doi 10. 2478/v.10122-012-0005-1

DOI: 10.2478/v10122-012-0005-1

# THE PROBLEM OF SYNTACTIC AMBIVALENCE IN CORPUS LINGUISTICS

Ireneusz Kida

Abstract: Ireneusz Kida. *The Problem of Syntactic Ambivalence in Corpus Linguistics*. Lingua Posnaniensis, vol. LIV (1)/2012. The Poznań Society for the Advancement of the Arts and Sciences. PL ISSN 0079-4740, ISBN 978-83-7654-103-7, pp. 57–63.

The purpose of this article is to present a technique of dual annotation of Old English ambivalent structures in diachronic annotated corpus linguistics. In languages there are often structures which are ambivalent, and it is difficult to establish whether they are main or dependent. These clauses are problematic for a corpus linguist annotating them for computer analysis of word order configurations. As a solution to this problem we suggest that such structures be annotated in two ways, namely on the one hand as main and on the other hand as dependent. Such a procedure allows one to obtain more objective results from word order analysis. Moreover, dual annotation is more flexible and is able to grasp the changeable nature of language.

Ireneusz Kida, Institute of English, Faculty of Philology, University of Silesia, ul. Gen. Stefana Grota-Roweckiego 5, PL – 41-205 Sosnowiec

## CORPUS LINGUISTICS

Corpus linguistics is a relatively young branch of linguistics but in recent years one can observe its increasing popularity.[1] As McEnry et al. (2006: 4) note, "nowadays, the corpus methodology enjoys widespread popularity. It has opened up or foregrounded many new areas of research [and] corpora have revolutionized nearly all branches of linguistics." The growing interest in corpus linguistics resulted in the construction of multiple diachronic and non-diachronic corpora for the analysis of various languages of the world. McEnry et al. (2006: 3) observe that "although the term corpus linguistics first appeared only in the early 1980s, corpus-based language study has a substantial history [and] the basic corpus methodology was widespread in linguistics in the early twentieth century." Moreover, they say that although linguists at that time did not use computers as a means of data storage, their methodology was essentially corpus-based in the sense that it was empirical and based on observed data. However, as they further observe, in late 1950s the corpus methodology was severy critisised and it became marginalised, but with the developments in computer technology the exploitation of massive corpora became possible, and the marriage of corpora with computer technology revived the interest in the corpus methodology.

[1] See McEnry and Wilson 2001.

## SOME ISSUES RELATED TO TEXT ANNOTATION

McEnery & Wilson (2001: 32) distinguish two kinds of corpora, namely, unannotated and annotated.[2] Unannotated corpora are characterised by being in their existing raw states of plain text, whereas annotated corpora are inhanced with various types of linguistic information and they are a very useful tool for a large scale analysis of different aspects of language. Since corpus linguistics is a relatively young field of study, the methodologies applied in the process of text annotation vary, and one cannot speak of any uniform and universal way of annotation of texts for electronic analyses. As Aarts & McMahon (2006: 44) observe, "corpus linguistics may be viewed as a methodology, but the methodological practices adopted by corpus linguists are not uniform."

## WHAT DEVELOPED FIRST IN INDO-EUROPEAN: PARATAXIS OR HYPOTAXIS?

The Proto-Indo-European language (i.e. PIE), according to Kiparsky (1995), was a paratactic language in which finite subordinate clauses were not embedded but adjoined, and this is confirmed by Sanskrit, Hittite, Old Latin and Classical Greek. When PIE split into different languages, most daughter languages, including Germanic, introduced an innovation in their syntax and departed a little from the original pattern.[3] As a result, dependent clauses became syntactically embedded in those languages and were taking up modifier or argument positions within the main clause. Lehmann (1974) also claims that PIE was paratactic. He maintains the view that it was an OV language and that the paratactic arrangement that is assumed for this language is typical of OV languages.[4] According to Delbrück (1900: 411–413; after Meier-Brügger et al. 2003: 245) "originally all sentences were coordinated alongside one another. […] The historical view, as it is generally accepted today, must have as its point of departure the hypothesis that there was a time at which there were only main clauses. […] The assertion that hypotaxis developed from parataxis has become the common heritage of the field." Quiles (2007: 237) says that "the oldest surviving texts consist largely of paratactic sentences, often with no connecting particles. New sentences may be introduced with particles, or relationship may be indicated with pronominal elements; but these are fewer than in subsequent texts." Furthermore, according to Meier-Brügger et. al. (2003) "along with parataxis (coordination), there is also evidence of hypotaxis (subordination) in Proto-Indo-European. […] The formal characteristics of subordinate clauses vary among the individual IE languages. In Proto-Indo-European, the accentuation of the finite verb is accepted as a formal characteristic of the subordinate clause as opposed to the main clause, in which the finite verb is not accentuated, except when it establishes the theme at the beginning of the sentence." The development of hypotaxis from parataxis seems to be not only typical of the Indo-European languages. Jucker (1991: 203) suggests that "it is gener-

[2] Curzan and Palmer (2006) use the terms unprincipled (or non-systematic) vs. principled corpora to mean unannotated and annotated corpora respectively.

[3] For more information on this issue see for example Friedrich 1975; Fortson 2004; Grace 1971; Greenberg 1963; Smith 1971.

[4] See also Lehmann 1972a, 1972b, 1973, 1992.

ally recognized that languages move from parataxis to hypotaxis. They do this on two levels. On the one hand, the proportion of hypotaxis versus parataxis tends to increase in the course of time, and, on the other hand, hypotactic constructions usually have paratactic origins."

According to Harris & Campbell (1995: 283–284) "the claim that hypotaxis develops from parataxis has often been made with reference to the first appearance of hypotaxis in a language, not to its repeated renewal. We use the term origin […] strictly to refer to the first appearance of a construction in a language; renewal refers to the continuing process of replacing or otherwise revising existing construction types." Moreover, they claim that when some authors write that hypotaxis developed out of parataxis, they "seem to have in mind conjunctionless joining, others loose joining, and still others discourse. Thus, even if, for the time being, we limit our inquiry to renewal, in approaching the question of whether hypotaxis develops out of parataxis we encounter the problem that different linguists have in mind different ideas of parataxis, and that at least some of them are vague." They add that there are basically two types of arguments that some authors use to support the view that it is parataxis that provides the source or prototype for hypotaxis. Namely, the first one relates to the ultimate origins of hypotaxis and "it is based on the claim that parataxis is more common in the early stage of a written language than is embedding." As regards the other type of argument, it "is based on the origin of the subordinator. Since subordinators in many languages originate as markers of questions – either yes/no or content questions – it is sometimes assumed that the subordinate clauses they mark must have originated as actual questions. Many languages have subordinators that originated as demonstrative pronouns and some investigators see this as evidence that those pronouns were 'pointing to' a loosely adjoined clause." However, they draw our attention to the fact that it does not necessarily have to be so because "it is by no means necessary to assume that the clause in which a particular innovative grammatical element is found developed out of the clause in which that grammatical element originated. It is logically possible that one *word* simply developed from another, with little reference to context. It is also possible that structural marking that developed in one context was later *extended* to another." They conclude that as a matter of fact the view that hypotaxis develops from parataxis and not vice versa is not supported by the evidence that comes from attested examples of the rise of the use of subordinators. According to Roberts (2007), the traditional and often repeated view that clausal subordination, or hypotaxis, is a relatively recent reanalysis of parataxis, or clause-chaining, should be abandoned although this view has a long history. He says that "the claim that earlier stages of certain languages may have lacked subordination altogether violates the uniformitarian hypothesis, the idea that all languages at all times reflect the same basic UG […] so I conclude that the traditional parataxis-to-hypotaxis idea should be abandoned, as it is conceptually problematic and in practice unrevealing" (p. 174).

Bednarczuk (1980: 145) observes that "the relation between parataxis and hypotaxis has not been precisely defined […] in spite of long discussions on the subject, which on the other hand allowed us to discover certain formal differences between them." Moreover, he claims that it is impossible to state empirically whether parataxis is older than hypotaxis or vice versa, or which of the two constructions has arisen from which. However, he notes that "the most widespread theory which says that hypotaxis has arisen from parataxis is based on the fact that it is less frequent in colloquial language and in children's speech, while in the historical development of different languages it expands at the cost of parataxis. [However,] in some languages, on the contrary, we can observe the expansion of parataxis at the cost of hypotaxis."

## AMBIVALENT NATURE OF SOME OLD ENGLISH CLAUSES: ARE THEY MAIN OR DEPENDENT?

No matter if hypotaxis develops from parataxis or hypotaxis gives rise to parataxis, it is logical to think that the development is not an abrupt one and that there is always a transition stage. JUCKER (1991: 203) says that “there must be one or possibly several intermediate stages between true parataxis and true hypotaxis and that there are constructions that are neither clearly paratactic nor clearly hypotactic but somewhere in-between. In most cases this development will have been not so much a matter of discrete steps, but rather a gradual movement, which makes it difficult to ascertain the exact status of a construction at any one time.”

Although Old English achieved quite an advanced stage of hypotaxis, we can often have problems with the classification of some clauses. As BAUGH & CABLE (1993: 66–67) indicate, “there are clear differences in our modern perceptions of Old English written in […] paratactic style and Old English written with many embedded clauses. The problem is in determining whether a particular clause is independent or subordinate, because the words that do the subordinating are often ambiguous. The Old English *þa* at the beginning of a clause can be either an adverb translated ‘then’ and indicating an independent clause, or a subordinating conjunction translated ‘when’ and introducing a dependent clause. Similarly, *þær* can be translated as ‘there’ or ‘where’, *þonne* as ‘then’ or ‘when’, *swa* as ‘so’ or ‘as’, *ær* as ‘formarly’ or ‘ere’, *siððan* as ‘afterward’ or ‘since’, *nu* as ‘now’ or ‘now that’, *þeah* as ‘nevertheless’ or ‘though’ and *forðam* as ‘therefore’ or ‘because.’” They also say that “in each pair the first word is an adverb, and the style that results from choosing it is a choppier style with shorter sentences, whereas the choice of the second word results in longer sentences with more embedded clauses.” Moreover, they note that “current research in Old English syntax aims to understand the use of these ambiguous subordinators and adverbs. The conclusions that emerge will affect our modern perception of the sophistication of Old English writing in verse and prose.” They also note that “we should be especially cautious about imposing modern notions that equate hypotaxis with sophistication and parataxis with primitiveness until we know more about the full range of syntactic possibilities in Old English. Ongoing research in this subject promises to revise our ideas of the grammatical, semantic, and rhythmic relationships in Old English verse and prose.”[5] Also MITCHELL (1985: §1879; after BAUGH & CABLE 1993: 67) warns us that it may be anachronistic to impose modern categories resulting from our translations into words like ‘then’ and ‘when’, “implying that the choice was simply between a subordinate clause and an independent clause in the modern sense of the words”. BAKER (2003: 29) observes that some linguists claim that Old English literature is generally characterised by parataxis, but it is not so, because it is only some Old English works, such as the *Anglo-Saxon Chronicle* for example, that tend to be paratactic, whereas other works, like King Alfred’s Preface to his translation of Gregory’s *Pastoral Care* for example, are characterised by hypotaxis. He further says that in Old English it can be difficult to tell independent clauses from subordinate clauses, and because of that it is a matter of some controversy how paratactic or hypotactic Old English was in fact.

[5] For more information on this issue see MITCHELL 1985, 1988; MITCHELL & ROBINSON 2007; BLAKE 1992; DENISON 1993; FISCHER et al. 2000; HOGG 1992; KOHONEN 1978; MOLENCKI 1997; PINTZUK 1993, 1995.

Dash (2005: 24–25) notes that ambiguity is very common especially at the lexical level in natural languages because a single lexical item may convey more than one sense, idea or event, depending on the context in which it is used. Moreover, he says that two types of ambiguity are found in a tagged corpus, namely, structural ambiguity and sequencial ambiguity. The former kind of ambiguity "is caused mostly for the non-inflected words where a root, due to its homographic structure, may belong to different lexical categories. It is also noted in case of some inflected words because root and suffix of these words are identical, although they belong to different lexical categories." As far as the latter kind of ambiguity is concerned, Dash claims that "it is mostly caused due to the presence of immediately following words, which when parsed together with the word under investigation, produces a meaning, which differs from their respective independent meanings." The prevalent existence of ambiguity in language poses a serious problem for the constructors of annotated corpora. Baker et al. (2006: 10) note that "in corpus annotation, in cases where there is a choice of two potential tags at one point in the text, it is not always possible to make a clear-cut decision. […] In some cases a portmanteau tag can be given in order to address the ambiguity. In other words, examining more of the surrounding context may help to solve the problem. However, in extremely ambiguous cases, the corpus builder may have to make a decision one way or the other. If this approach is taken then the decision would need at least to be applied with consistency throughout the corpus. In general, decisions regarding ambiguous cases should be covered in the documentation that comes with the corpus." Pala et al. (1997: 523) say that the most reasonable way of building large annotated corpora is via an automatic tagging of the texts by means of computer programmes. However, they add that "natural languages display rather complex clause and therefore it is no surprise that the attempts to process them by the simple deterministic algorithms do not always yield satisfactory results. The result is that the present tagging programmes are not able to give fully reliable results and there are many ambiguities in their output."

## THE WAY WE SEE PARA-HYPOTAXIS

We employ the term *para-hypotaxis* to mean something different from what it is usually used to mean. The term usually means a situation in which one is dealing with subordinate clause syntax with only a coordinate interpretation possible, or in which the writer treats as coordinate, clauses which would appear to require subordination (see for example Scaglione 1972). According to Mazzoleni (2002) the term 'parahypotaxis' is the name traditionally assigned to Old Italian sequences of dependent clauses with following main clauses introduced by *e* 'and', *sì* 'thus'; he also suggests that the conjunction *ma* 'but' should be taken into account here too. Van Valin (2005: 187) discusses the problem of switch-reference constructions in Amele, Kewa and Chuave that are examples of neither subordination nor simple coordination. He says that "these constructions are therefore a kind of dependent coordination, in which units of equivalent size are joined together in a coordinate-like manner relation but share some grammatical category, e.g. tense or mood." He also says that this linkage or nexus relation was termed 'cosubordination' in Olson (1981). We personally use the term *para-hypotaxis* with respect to clauses that have an ambivalent status. We call them *ambivalent para-hypotactic clauses* or *PH clauses*, and they belong to the so called *para-hypotaxis*. These clauses are ambivalent because on the one hand they seem to behave like main clauses and on the other hand they seem to behave like dependent clauses. This fact means that they can be analysed in two ways, namely either as main clauses being in paratactic relation to the immediately preceding/following clauses, or as dependent clauses being in hypotactic relation to the clauses immediately preceding/following them. These clauses pose a problem to a linguist dealing with their annotation, because whenever he comes across them he has to make a subjective decision as to how to approach them.[6] To our knowledge, in annotated corpus linguistics when clauses are annotated for the analysis of word order configurations the common trend is to annotate them only in one way, that is, either the way that they are treated as main or the way that they are treated as dependent. The result of such an approach is that what was annotated in a given way has to stay in this form, because once the linguistic material was annotated rigidly, it is not possible to analyse it from a different perspective. In other words, the annotated corpora that are produced for the analysis of Old English texts, as well as texts written in other languages, are not flexible and do not reflect the dynamic and changeable nature of language, and therefore they do not allow one to grasp *the ambivalent* during the analysis. There is a danger in such a rigid approach in the sense that the end-users, especially the unexperienced ones, who would like to make use of such corpora for word order analysis, will usually take it for granted that the annotated corpora that they are making use of were annotated in the right way and that the results obtained from the analysis are objective and cannot be questioned. However, if one does not approach such rigid corpora with some distance, one will run the risk of obtaining data that are inherently wrong, at least in some respect. Therefore, there is a strong need for an adequate approach to the linguistic material in the annotation process. Therefore we suggest that ambivalent clauses be annotated in different ways within the same corpus, namely, on the one hand as if they were main and, on the other, as if they were dependent. Such annotation is more capable of taking into account and reflecting dynamic and changeable character of language in the analysis of word order configurations, than rigid annotation.

## REFERENCES

Aarts Bas, McMahon April. 2006. *The Handbook of English Linguistics*. Oxford: Blackwell Publishing.

Baker Peter S. 2003. *Introduction to Old English*. Oxford: Blackwell Publishing.

Baker Paul, Hardie Andrew, McEnry Tony. 2006. *A Glossary of Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Battye Adrian, Roberts Ian (eds). 1995. *Clause Structure and Language Change*. New York: Oxford University Press.

Bednarczuk Leszek. 1980. “Origin of Indo-European Parataxis.” In: Ramat 1980: 145–154.

Blake Norman (ed.). 1992. *The Cambridge History of the English Language*. Vol. 2. Cambridge: CUP.

Curzan Anne, Palmer Chris. 2006. “The Importance of Historical Corpora, Reliability and Reading. In: Facchinetti & Rissanen 2006: 17–34.

Delbrück Berthold. 1900. *Vergleichende Syntax der indogermanischen Sprachen*. Vol. 3. New York: Cambridge University Press.

Denison David. 1993: *English Historical Syntax: Verbal Constructions*. London–New York: Longman.

Facchinetti Roberta, Rissanen Matti (eds). 2006. *Corpus-based Studies of Diachronic English*. Bern: Peter Lang.

Fischer Olga, van Kemenade Ans, Koopman Willem, van der Wurff Wim. 2000. *The Syntax of Early English*. Cambridge: Cambridge University Press.

[6] Cf. Baker et al. (2006).

Friedrich Paul. 1975. *Proto-Indo-European Syntax*. Journal of Indo-European Studies. Monograph 1.

Fortson Benjamin W. 2004. *Indo-European Language and Culture*. Oxford: Blackwell Publishing.

Greenberg Joseph H. 1963. "Some Universals of Language with Special Reference to the Order of Meaningful Elements." In: Greenberg (ed.) 1963: 73–113.

Greenberg Joseph H. (ed.). 1963. *Universals of Language*. Cambridge, MA: MIT Press.

Harris Alice, Campbell Lyle. 1995. *Historical Syntax in Cross-linguistic Perspective*. New York: Cambridge University Press.

Hogg Richard M. (ed.). 1992. *The Cambridge History of the English Language*. Vol. 1. Cambridge: Cambridge University Press.

Jucker Andreas H. 1991. "Between Hypotaxis and Parataxis. Clauses of Reason in Ancrene Wisse." In: Kastovsky 1991: 203–219.

Kastovsky Dieter (ed.). 1991. *Historical English Syntax*. Berlin: Mouton de Gruyter.

Kiparsky Paul. 1995. "Indo-European Origins of Germanic Syntax." In: Battye & Roberts 1995: 140–169.

Kohonen Viljo. 1978. *On the Development of English Word Order in Religious Prose Around 1000 and 1200 A.D.* Åbo: Research Institute of the Åbo Akademi Foundation.

Lehmann Winfred P. 1974. *Proto-Indo-European Syntax*. Austin–London: University of Texas Press.

Mazzoleni Marco. 2002. "La 'paraipotassi' con *ma* in italiano antico: verso una tipologia sintattica della correlazione." *Verbum 4*. Budapest, Akadémiai Kiadó: 399–427.

McEnry Tony, Wilson Andrew. 2001. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

McEnery Tony, Xiao Richard, Tono Yukio. 2006. *Corpus-based Language Studies: An Advanced Resource Book*. New York: Routledge.

Meier-Brügger Michael, Fritz Matthias, Mayrhofer Manfred. 2003. *Indo-European Linguistics*. Berlin: Walter de Gruyter.

Mitchell Bruce. 1985. *Old English Syntax*. 2 vols. Oxford: Clarendon Press.

Mitchell Bruce. 1988. *On Old English*. New York: Basil Blackwell.

Mitchell Bruce, Robinson Fred C. 2007. *A Guide to Old English*. Oxford: Blackwell Publishing.

Molencki Rafał. 1997. "Surface Word Order in Subordinate Clauses in Early West Saxon Prose. *Linguistica Silesiana* 18, 29–38.

Olson Michael. 1981. *Barai Clause Junctures: Towards a Functional Theory of Inter-clause Relations*. Ph.D. Dissertation Australian National University.

Pala Karel, Rychlý Pavel, Smrž Pavel. 1997. "DESAM – Annotated Corpus for Czech." In: Plášil & Jeffrey 1997: 523–530.

Pintzuk Susan. 1993. *Phrase Structure Variation in Old English*. Paper presented at the Second Diachronic Generative syntax Workshop at University of Pennsylvania, 5–8 November 1992. MS.

Pintzuk Susan. 1995. "Variation and Change in Old English Clause Structure." *Language Variation and Change* 7, 229–260.

Plášil Frantisek, Jeffrey Keith G. (eds). 1997. *SOFSEM'97: Theory and Practice of Informatics*. Berlin: Springer.

Ramat Paolo (ed.). 1980. *Linguistic Reconstruction and Indo-European Syntax*. Amsterdam: John Benjamins.

Roberts Ian. 2007. *Diachronic Syntax*. New York: Oxford University Press.

Scaglione Aldo. 1972. *The Classical Theory of Composition. From Its Origins to the Present: a Historical Survey*. Chapel Hill: University of North Carolina Press.

Smith Jesse. 1971. *Word Order in the Older Germanic Dialects*. University of Illinois dissertation. Urbana-Champaign.

Van Valin Robert D. 2005. *Exploring The Syntax-semantics Interface*. Cambridge: Cambridge University Press.

*Allatum die 30 mensis Januarii anno 2012*