

**Title:** Introduction to corpus linguistics

**Author:** Ireneusz Kida

**Citation style:** Kida Ireneusz. (2013). Introduction to corpus linguistics. "Linguistica Silesiana" (Vol. 34 (2013), s. 133-144).

Linguistica Silesiana 34, 2013
ISSN 0208-4228

IRENEUSZ KIDA
University of Silesia

# INTRODUCTION TO CORPUS LINGUISTICS

The paper aims at presenting selected notes and remarks on corpora and corpus linguistics. It starts with a brief history of corpus linguistics. It occurs that although corpus linguistics is a relatively young branch of linguistics, it managed to revolutionise all branches of linguistics. Afterwards, the notion of *corpus* and different types of corpora are discussed . In general, we can say that, on the one hand, there are annotated and unannotated corpora, and, on the other, diachronic and synchronic ones. In the following sections of the article the notions of corpus composition, annotation, size and representativeness are discussed, and towards the end of the paper a list of the advantages of corpus linguistics is presented and some further conclusions drawn.

## 1. Corpus linguistics

Corpus linguistics is a relatively young branch of linguistics and Lindquist (2009: 1) defines it as "a methodology, comprising a large number of related methods, which can be used by scholars of many different theoretical leanings." However, Aarts and McMahon (2006: 44) observe that "corpus linguistics may be viewed as a methodology, but the methodological practices adopted by corpus linguists are not uniform." McEnry et al. (2006: 3) note 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 severely criticised 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.

McEnry and Wilson (2001: 24) also note that "although the methodology went through a period of relative neglect for two decades, it was far from abandoned. Indeed, during this time essential advances in the use of corpora were made. Most importantly of all, the linking of the corpus to the computer was completed during this era. Following these advances, corpus studies boomed from 1980s onwards, as corpora, techniques and new arguments in favour of the use of corpora became more apparent." This boom, they say, continues currently and corpus linguistics is becoming more and more mature methodologically, and the range of languages that are addressed by corpus linguists is growing annually. Lindquist (2009) notes that the first electronic collection of English texts to be used for linguistic research, was compiled by the pioneers in corpus linguistics Nelson Francis and Henry Kučera in the early 1960's at Brown University, US. This electronic collection of English texts is referred to as the Brown Corpus, and it is regarded as the first non-diachronic computer corpus ever developed. Soon after computers started becoming more and more powerful, which caused that the field of corpus linguistics was developing faster and faster. It gained a phenomenal momentum in the 2000s, and in recent years one can observe that it is more and more popular, not only among scholars[1].

## 2. Definition of a *corpus*

In the past, as Lindquist (2009) observes, the word *corpus* (Lat. 'body') was used about the total works written by an individual author or a certain mass of texts, as for example "The Shakespeare corpus". These were the so-called pre-electronic corpora. Nowadays, the term *corpus* is almost always associated with *electronic corpus,* which is a collection of texts stored on some kind of digital medium to be used by linguists with the purpose of retrieving linguistic items for research or by lexicographers in making dictionaries. According to Renouf (1987) the term 'corpus' refers to a collection of written or spoken texts which is stored and processed on computer for the purposes of linguistic research. Sinclair (1991: 171) says that "a corpus is a collection of naturally-occurring language text, chosen to characterize a state or variety of a language. In modern computational linguistics, a corpus typically contains many millions of words: this is because it is recognized that the creativity of natural language leads to such immense variety of expression that it is difficult to isolate the recurrent patterns that are the clues to the lexical structure of the language."

Sinclair (1991) distinguishes two types of corpora, namely sample corpus and monitor corpus. The former is a finite collection of texts, often chosen with great care and studied carefully. On establishing a sample corpus, one cannot add anything to it or change in any way. As for the latter, it is a continually-growing one

[1] See McEnry and Wilson (2001).

and it re-uses language text which has been prepared in machine-readable form for other purposes, like for typesetters of newspapers, magazines, books and also word-processors; and the spoken language basically for legal and bureaucratic reasons. McEnery and Wilson (2001: 32) also 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 enhanced with various types of linguistic information and they are a very useful tool for a large scale analysis of different aspects of language. Some of the most common types of corpus annotation are textual mark-up, part-of-speech (POS) tagging, syntactic annotation (parsing), semantic annotation, prosodic annotation, pragmatic annotation, discourse annotation, phonetic annotation and stylistic annotation (Leech 2004).

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. However, Leech (2004) notes that more recently there has been a far-reaching trend to standardise the representation of all phenomena of a corpus, including annotations, by means of a standard mark-up language – usually one of the series of related languages SGML, HTML, and XML. One of the advantages of using these languages for encoding features in a text is that they allow the interchange of documents, including corpora, between one user, or research site, and another. In this sense, Leech says, SGML/HTML/XML have developed into a world-wide standard which can be applied to any language, both spoken and written, as well as to languages of different historical periods.

Finally, Nesselhauf (2011) distinguishes the following kinds of corpora: general/reference corpora which aim at representing a language or a language variety as a whole and they contain both spoken and written language (e.g. The British National Corpus or The Bank of English), historical corpora (vs. corpora of present-day language) which aim at representing an earlier stage or earlier stages of a language (e.g. The Helsinki Corpus or the ACHER), regional corpora which aim at representing one regional variety of a language (e.g. The Wellington Corpus of Written New Zealand English), learner corpora (vs. native speaker corpora) which aim at representing the language as produced by learners of this language (e.g. The International Corpus of Learner English), multilingual corpora (vs. one-language corpora) which aim at representing several, at least two, different languages, often with the same text types to enable contrastive analysis (e.g. The PROIEL Corpus, a parallel corpus of New Testament texts from different languages like Greek, Latin, Gothic, Old Church Slavonic and Classical Armenian), and spoken corpora (vs. written vs. mixed corpora) which

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

aim at representing spoken language (e.g. The London-Lund Corpus of Spoken English).[3]

## 3. Corpus composition

Sinclair (2005) discusses some instructions that should be followed in the composition of a corpus and in the compilation of language samples. Below are the ten principles that he considers as fundamental:

1. The contents of a corpus should be selected without regard for the language they contain, but according to their communicative function in the community in which they arise.
2. Corpus compilers should strive to make their corpus as representative as possible of the language from which it is chosen.
3. Only those components of corpora which have been designed to be independently contrastive should be contrasted.
4. Criteria for determining the structure of a corpus should be small in number, clearly separate from each other, and efficient as a group in delineating a corpus that is representative of the language or variety under examination.
5. Any information about a text other than the alphanumeric string of its words and punctuation should be stored separately from the plain text and merged when required in applications.
6. Samples of language for a corpus should, wherever possible, consist of entire documents or transcriptions of complete speech events, or should get as close to this target as possible. This means that samples will differ substantially in size.
7. The design and composition of a corpus should be documented fully with information about the contents and arguments in justification of the decisions taken.
8. The corpus compiler should retain, as target notions, representativeness and balance. While these are not precisely definable and attainable goals, they must be used to guide the design of a corpus and the selection of its components.
9. Any control of subject matter in a corpus should be imposed by the use of external, and not internal, criteria.
10. A corpus should aim for homogeneity in its components while maintaining adequate coverage, and rogue texts should be avoided.

---

[3] More information on the different types of corpora and corpus linguistics can be found in Rissanen (2012), Facchinetti and Rissanen (2006), Haug et al. (2009), Kennedy (1998), Meyer (2002), Rizzo (2010), Lewandowska-Tomaszczyk (2007), Bermudez-Otero et al. (2000), as well as at http://www.helsinki.fi/varieng

## 4. Corpus annotation

As far as annotation is concerned, McEnry et al. (2006: 33) say that "corpus annotation can be achieved fully automatically, by a semi-automatic interaction between human being and the machine, or entirely manually by human analysts". They also say that the annotation of a corpus may have many forms and it can be undertaken at different levels:

1. At the phonological level; where corpora can be annotated for syllable boundaries (phonetic/phonemic annotation) or prosodic features (prosodic annotation).
2. At the morphological level; where corpora can be annotated in terms of prefixes, stems and suffixes (morphological annotation).
3. At the lexical level; where corpora can be annotated for parts of speech (POS tagging), lemmas (lemmatisation), and semantic fields (semantic annotation).
4. At the syntactic level; where corpora can be annotated to show anaphoric relations (coreference annotation), pragmatic information like speech acts (pragmatic annotation) or stylistic features such as speech and thought presentation (stylistic annotation).

They observe that out of the different types of annotation POS tagging is the most widespread type of annotation, and that syntactic parsing is also developing quite fast. However, such types of annotation as discoursal annotation and pragmatic annotation are presently relatively underdeveloped. Květoň and Oliva (2002:19) observe that "the quality of corpus annotation is certainly among the pressing problems in current corpus linguistics. This quality, however, is a many-faceted problem in itself, comprising both issues of a rather theoretical nature and also quite practical matters"[4].

In the light of the lack of uniform annotated corpora and the different research needs that the researchers have with respect to language, Květoň and Oliva (2002: 255) encourage linguists to write their own computer programs for the construction of their own corpora. They say that there are several concrete advantages to writing one's own computer programs rather than relying on available concordancing software. Below I enumerate the arguments that they provide to support their claim:

1. Perhaps most importantly, writing programs allows one to conduct analyses that are not possible with concordances.
2. One can do many analyses more quickly and accurately.
3. One can tailor the output of the analysis to fit one's research needs.
4. When one writes one's own programs, there is no limit to the size of the corpus that can be analysed.

---

[4] See also Sinclair (2004b).

Generally speaking, writing one's own computer program for the construction of one's own corpus and its analysis "opens up a wider variety of options in the research questions that you can investigate" (Květoň and Oliva 2002: 256).

I could also provide my own arguments in favour of the idea of creating one's own corpus that result from our experience[5] with diachronic annotated corpus linguistics:

1. One can freely choose texts to be annotated. This means that one can even analyse texts, or fragments of texts, that have not yet been annotated by anyone.
2. One can make the annotation simpler and more user-friendly.
3. One does not feel the limits and imperfections imposed on one by the existing annotated corpora.
4. One can find one's own solutions to the problems that might occur during the annotation process.
5. One can make use of simple and widely accessible computer programs for the construction of one's annotated corpora; for example the Microsoft Word.
6. One can construct one's own corpus the way that it can be modified and adapted to the needs of one's present and future research.
7. One can construct a corpus that will reflect one's own ideas about language
8. One can construct a corpus that will allow one a dual/multiple approach to ambiguous/ambivalent structures in language instead of a rigid one-way approach.

Furthermore, McEnery and Wilson (2001: 32) observe that "unannotated corpora have been, and are, of considerable use in language study, but the utility of the corpus is considerably increased by the provision of annotation. The important point to grasp about an annotated corpus is that it is no longer simply a body of text in which the linguistic information is implicitly present. [Moreover,] a corpus, when annotated, may be considered to be a repository of linguistic information, because the information which was implicit in the plain text has been made explicit through concrete annotation". According to Leech (1993; after Dash 2005: 5) there are seven maxims that should be applied strictly in the annotation of texts:

1. It should always be easy to dispense with annotation, and revert to the raw corpus. The raw corpus should be recoverable.
2. The annotations should correspondingly be extractable from the raw corpus, to be stored independently, or stored in an interlinear format.
3. The scheme of analysis presupposed by the annotations – the annotation scheme – should be based on principles or guidelines accessible to the user.

---

[5] Cf. Kida (2007a, 2007b, 2009, 2010a, 2010b, 2011 and 2012).

4. It should be made clear beforehand about how and by whom all the annotations were applied.
5. The user must be made aware that the annotation applied in the corpus is not infallible, but simply a potentially useful tool.
6. Annotation schemes should preferably be based as far as possible on 'consensual', theory-neutral analyses of the corpus data.
7. No one annotation scheme can claim authority as a standard, although as a matter of fact interchange 'standards' may arise, through widening availability of annotated corpora, and perhaps should be encouraged.

Dash (2005) observes that in annotated corpus linguistics there are basically three important criteria that are usually considered as important in any kind of annotation. These criteria are: consistency, accuracy and speed. Firstly, as regards consistency, it concerns the uniformity in annotation throughout the whole text of a corpus. Secondly, accuracy is about the freedom from any kind of error in the tagging to adhere to the definitions and guidelines concerning the scheme of annotation. Thirdly, the automatic implementation of the scheme of annotation should be possible on a very large data quantity within a very short span of time.

## 5. Corpus size and reprsentativeness

Above I mentioned the problem of the lack of uniformity in annotated corpus linguistics. However, it is not the only problem that corpus linguists are facing. Among others, there is also the problem of how representative a given corpus is, and the problem of what size it should have in order to be representative. Kohnen (2007) notes that a first major difficulty in corpus linguistics is connected with corpus size as it is not known exactly how large corpora must be in order to qualify for valid linguistic research. Moreover, He says that on surveying the field one can get the impression that even in the age of so-called second-generation mega corpora, researchers seem to be less confident about the 'definite' size that corpora should have. Kohnen also notes that the problem of representativeness is another central concern in corpus linguistics and corpus linguists should aim at building such corpora that would be representative. However, he admits that when we are dealing with representativeness, many researchers are very reserved.

According to Biber et al. (1998), a corpus is not a mere collection of texts. A corpus should rather seek to represent a language or some part of language. Therefore the appropriate design for a corpus is dependent upon what it is going to represent and the kinds of research questions that can be addressed, and the generalisability of the results of the research, in turn, is determined by the representativeness of the corpus. They conclude that "it is important to realize up front that representing a language – or even part of a language – is a problematic task. We do not know the full extent of variation in languages or all the contextual variables that need to be covered in order to capture all

variation in texts" (p. 246). Mukherjee (2004) admits pessimistically that it is not possibile to attain absolute representativeness. Furthermore, according to Römer (2005: 41), "a large corpus can generally be regarded more representative of the type of language it consists of than a small corpus which contains the same kind of language. Of course, any small corpus is better than no corpus at all, but if the choice is between a small and a large corpus of the same (or similar) kind of material, I would always go for the latter." Leech (2007: 138) observes that "there is one rule of thumb that few are likely to dissent from. It is that in general, the larger a corpus is, and the more diverse it is in terms of genres and other language varieties, the more balanced and representative it will be." According to Leech (1991: 27) we can say that a corpus is representative when "the findings based on its contents can be generalized to a larger hypothetical corpus." He moreover observes that the issue of corpus representativeness must be considered largely as an act of faith because at present there is no way of ensuring it or evaluating it in an objective way, although a great deal of research is carried out with respect to this issue[6].

## 6. Corpus application

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 revolutionised nearly all branches of linguistics." Ezquerra and Hurtado (1996: 41; after Endres and Wagner 1992) mention the following disciplines in which corpora find their application:

- Theoretical linguistics: traditional linguistics disciplines such as syntax, morphology, phonetics, etc.
- Lexicology and lexicography.
- Computational linguistics and related fields: language processing, computer analysis, language recognition, speech synthesis, information sciences, knowledge acquisition, expert systems, automated translation, text processing, language statistics, etc.
- Theory and practice of communication, including publishing.
- Psycholinguistics and related fields: neuropsychology, language philosophy, discourse analysis, text linguistics, etc.
- Computer assisted teaching: learning, stylistics, orthography, etc.

Lindquist (2009) notes that compiling corpora can be very time-consuming and expensive, therefore there must be considerable gains for the linguists to

---

[6] For more discussion concerning this issue, as well as the issue of size, authenticity, sampling, etc. see for example Tognini-Bonelli (2001), Sánchez et al. (1995), McEnry and Wilson (2001), Sinclair (2005), and Wynne (2005).

justify the effort. He says that the major advantages of corpora are speed and reliability, as by using a corpus the linguist can investigate more material than in manual investigation, and within a shorter time too, and moreover he can obtain more exact results. Lindquist (2009:9) also presents a list of the advantages of corpus linguistis that can be found in Svartvik (1992: 8-10), one of the founding fathers of ICAME, the International Computer Archive of Modern and Medieval English, that was started in 1977 in Oslo. I will mention only some of the advantages given:

- Corpus data are more objective than data based on introspection.
- Researchers can share the same corpus data instead of always compilig their own.
- Corpora provide the possibility of total accountability of linguistic features.
- Computerised corpora give researchers all over the world access to the data.
- Corpus data are ideal for non-native speakers of the language.
- Corpus data are excellent for studies of language variation.
- Corpus data provide frequency of occurrence of linguistic items.
- Corpus data give essential information for a number of applied areas, like language teaching and language technology (machine translation, speech synthesis, etc.)

These and many other advantages of corpora over manual investigation not mentioned here are the reasons for the fact that corpora are constantly being developed and there is a growing interest in corpus linguistics.

## 7. Conclusion

Corpus linguistics and corpora have been with us starting from the early 1960's, when linguistic research for the first time started to be assisted by means of computers. The first computer corpus, namely the Brown Corpus, was compiled by the pioneers in corpus linguistics Nelson Francis and Henry Kučera at Brown University, US. Since then on more and more corpora started to appear. However, in the twenty first century there has been an unprecedented boom in the compilation of corpora all over the world. Although languages like English and Spanish, and some other major languages of the world, have been enjoying most attention on part of corpus linguists, corpora have also been constructed for minor languages, like Polish, Czech, Croatian, Greek, and many other. There are different types of corpora. The earliest corpora were basically unannotated text corpora consisting only of pure texts, or samples of texts, but afterwards more and more annotated corpora began to appear. Annotated corpora are designed not so much for the analysis of pure texts, as for the analysis of text structure involving the analysis of word order, parts of speech, grammatical and phonological structure of words etc. However, annotated corpora are much more dif-

ficult to construct, are more challenging and usually involve many language experts, much financial means and are more time-consuming. This is the reason why there are fewer annotated corpora than unannotated ones, and why they are usually smaller. Nevertheless, in recent years annotated corpora are proliferating and soon, due to technological advances, the speed at which they are constructed might be equal to the speed at which uannotated corpora are. Both annotated and unannotated corpora have greatly facilitated the work of linguists working in practically all areas of linguistics, and fields related to linguistics and not only. Owing to the existence of corpora, which nowadays are becoming bigger and bigger, and the most recent ones contain many millions of words, it is possible to do a large-scale research, which leads to more objective results, than a small-scale research, based on a small number of texts or text samples, which is now a thing of the past.

## References

Aarts, B. and A. McMahon 2006. *The handbook of English linguistics*. Oxford: Blackwell Publishing Ltd.

Bermúdez-Otero, R. et al. (eds) 2000. *Generative theory and corpus studies: a dialogue from 10ICEHL*. Berlin: Mouton de Gruyter.

Biber, D., S. Conrad and R. Reppen 1998. *Corpus linguistics. Investigating language clause and use*. New York: Cambridge University Press.

Curzan, A. and C. Palmer 2006. The importance of historical corpora, reliability and reading. In R. Facchinetti and M. Rissanen (eds) *Corpus-based studies of diachronic English*, 17-34. Bern: Peter Lang.

Dash, N.S. 2005. *Corpus linguistics and language technology*. New Delhi: Krishan Mittal for Mittal Publications.

Endres, B. and F. Wagner 1992. Synoptic report on the needs of corpus users. In *Technical Report (II) prepared in the Institut für Deutsche Sprache, Mannheim, december 1992, for the WP2 of the project NERC, number 119.*

Ezquerra, M.A. and L.L. Hurtado 1996. Las industrias de la lengua y las aplicaciones de los corpora. In *Scripta philological in memoriam Manuel Taboada Cid.* A Coruña: Servicio de publicaciones, Universidade da Coruña.

Facchinetti, R. and M. Rissanen (eds). 2006. *Corpus-based studies of diachronic English*. Bern: Peter Lang.

Haug, D.T. et al. 2009. Computational and linguistic issues in designing a syntactically annotated parallel corpus of Indo-European languages. *TAL* 50(2): 17-45.

Kennedy, G. 1998. *An introduction to corpus linguistics.* London: Longman.

Kida, I. 2007a. The construction of a tagged corpus and the investigation of the change from OV to VO in English. *Academic Papers of College of Foreign Languages* 4: 77-86. Wydawnictwo Wyższej Szkoły Lingwistycznej, Częstochowa.

Kida, I. 2007b. *Word order tendencies in Mediaeval English against the Indo-European background*. Katowice: Wydawnictwo Uniwersytetu Śląskiego.

Kida, I. 2009. Syntactic differences between Gothic and Greek in Wulfila's translation of the Bible. In *Studia językoznawcze dedykowane Profesorowi Kazimierzowi Polańskiemu. W kręgu teorii*, 117-124. Katowice: Wydawnictwo Uniwersytetu Śląskiego.

Kida, I. 2010a. The emergence of subordinate (hypotactic) clauses in Old English from paratactically conjoined structures. *Linguistica Silesiana* 31: 115-122.

Kida, I. 2010b. Construction of an annotated corpus of Polish mediaeval texts for the analysis of word order configurations. *Research on Slavic Languages 14-2*: 157-166. Korean Association of Slavic Linguistics (KASL).

Kida, I. 2011. Hebrew, Latin and Greek word order of the Genesis in comparison. *Zeszyty naukowo-dydaktyczne. Nauczycielskie Kolegium Języków Obcych w Zabrzu*, 21-31.

Kida, I. 2012. Problem syntaktycznej ambiwalencji w anotowanym językoznawstwie korpusowym. *Linguistica Posnaniensia* 54(1): 57–63. ISSN (Online) 2083-6090, ISSN (Print) 0079-4740, DOI: 10.2478/v10122-012-0005-1.

Kohnen, Th. 2007. From Helsinki through the centuries: the design and development of English diachronic corpora. In P. Pahta, I. Taavitsainen, T. Nevalainen, and J. Tyrkkö, (eds) *Studies in variation, contacts and change in English volume 2. Towards multimedia in corpus studies*. Research Unit for Variation, Contacts and Change in English (VARIENG), University of Helsinki. http://www.helsinki.fi/varieng/journal/volumes/02/kohnen/

Květoň, P. and K. Oliva 2002. Achieving an almost correct PoS-tagged corpus. In *Text, speech and dialogue*. 5th international congress, TSD 2002, 19-26. Berlin: Springer.

Leech, G. 1991. The state of the art in corpus linguistics. In K. Aijmer and B. Altenberg (eds) *English corpus linguistics: studies in honour of Jan Svartvik*. London: Longman.

Leech, G. 1993. Corpus annotation schemes. *Literary and Linguistic Computing* 8: 275-81.

Leech, G. 2004. Adding linguistic annotation. In M. Wynne (ed.) *Developing linguistic corpora: a guide to good practice.* Available at: http://www.ahds.ac.uk/guides/linguistic-corpora/index.htm

Leech, G. 2007. New resources, or just better old ones? The Holy Grail of representativeness. In M. Hundt, N. Nesselhauf and C. Biewer (eds.). *Corpus linguistics and the web*, 133-150. Amsterdam: Editions Rodopi B.V.

Lewandowska-Tomaszczyk, B. (ed.) 2007. Corpus linguistics, computer tools, and applications – state of the art: PAIC 2007 (Lodz Studies in Language). Peter Lang.

Lindquist, H. 2009. *Corpus linguistics and the description of English.* Edinburgh: Edinburgh University Press Ltd.

McEnery, T., R. Xiao and Y. Tono 2006. *Corpus-based language studies: an advanced resource book*. New York: Routledge.

McEnry, T. and A. Wilson 2001. *Corpus linguistics*. Edinburgh: Edinburgh University Press Ltd.

Meyer, Ch.F. 2002. *English corpus linguistics*. *An introduction*. Cambridge: Cambridge University Press.

Mukherjee, J. 2004. The state of the art in corpus linguistics: three book-length perspectives. *English Language and Linguistics* 8(1):103-119.

Nesselhauf, N. 2011. *Corpus linguistics: a practical introduction*. An electronic publication, accessible at: http://www.as.uni-heidelberg.de/personen/Nesselhauf/files/Corpus%20Linguistics%20Practical%20Introduction.pdf

Renouf, A. 1987. *Corpus development.* In J. Sinclair (ed.) *Looking up*. London: Collins.

Rissanen, M. 2012. Corpora and the study of English historical syntax. In M. Kytö (ed.) *English corpus linguistics: crossing paths*, 197-220. Amsterdam: Rodopi.

Rizzo, C.R. 2010. Getting on with corpus compilation: from theory to practice. *ESP World*, Issue 1(27), Vol. 9. Available at: *http://www.esp-world.info.*

Römer, U. 2005. *Progressives, patterns, pedagogy: a corpus-driven approach to English progressive forms, functions, contexts and didactics*. Amsterdam: John Benjamins B.V.

Sánchez, A., P. Cantos, R. Sarmiento and J. Simón 1995. *Corpus lingüístico del español contemporáneo. Fundamentos, metodología y análisis*. Madrid: SGEL.

Sinclair, J. 1991. *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Sinclair, J. 2004b. Intuition and annotation – the discussion continues. In K. Aijmer and B. Altenberg (eds). *Advances in corpus linguistics. Papers from the 23rd International Conference on English Language Research on Computerized corpora (ICAME 23)*, 39-59. Amsterdam/New York: Rodopi.

Sinclair, J. 2005. Corpus and text – basic principles. In M. Wynne (ed.) *Developing linguistic corpora: a guide to good practice*. Oxford: Oxbow Books. Available at: *http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm*

Svartvik, J. (ed.) 1992. *Directions in corpus linguistics*. Berlin: de Gruyter.

Tognini-Bonelli, E. 2001. *Corpus linguistics at work*. Amsterdam: John Benjamins.

Wynne, M. (ed.). 2005. *Developing linguistic corpora: a guide to good practice*. Oxford: Oxbow Books. Available at: *http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm*

**Websites**:

http://www.helsinki.fi/varieng