Text by Dr. Swantje Westpfahl

Digital Humanities

Digital Humanities (often abbreviated DH) is an area of science, research and teaching at the intersection between humanities and computing. Digital Humanities can be applied to a variety of subjects, e.g. art history, cultural, musical, or literature studies, history, as well as anthropology, archeology and many more. As the name already implies, the discipline combines two fields of research. On the one hand computer-assisted methods, e.g. data-mining, statistics, visualization, ontologies and information management, are either used to answer existing research questions in the humanities, or to question existing theoretical paradigms, e.g. to verify or falsify theories, or to open up new access points for research questions and analysis. On the other hand, traditional methods and approaches of the humanities are used to deploy them on digital media, texts or objects, i.e. the digital world becomes the research object of the humanities. In linguistics, digital humanities play a role as well - for example in the construction and exploitation of language corpora in order to answer linguistic research questions.

What is a corpus?

Generally, a text corpus is a digital collection of written or transcribed spoken texts. Text corpora exist in various languages and are built for various purposes: mainly for research in linguistics, but also in the disciplines of e.g. literature studies, history or law. As the work with corpora in linguistics became more and more important, a new discipline has evolved, namely corpus linguistics. Corpus linguists build and use the corpora e.g. for validating linguistic rules through checking for attestations or to do statistical analysis and hypothesis testing on a large amount of data. Some corpora are especially built and designed in order to answer specific research questions e.g. corpora of children's language, dialect or complete editions of an author's works. Others aim at representing one language in its entirety, i.e. through a representative variety of text types which allow deductions about the system of the language. They are called reference corpora. Examples for reference corpora for the English language are e.g. the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA).

Q&A - Working with a corpus

Using corpora in diachronic linguistics, some thoughts in advance on basic features of corpus linguistics might be helpful. In order to find the right corpus to answer your research question there are some questions you might want to answer first:

1. Where do I find corpora and how can I access them?

To find corpora which are useful in the field of diachronic linguistics one can consult the corpus finder of the **Corpus Resource Database** (CoRD). There, an overview on various corpora of the English language is given; e.g. corpora with texts from Old English, Middle English, Early Modern English and Present-Day English. Additionally, you will also find information on whether you can use the corpora online, whether you need to download them, or whether you might need to pay for a license to use them.

2. Which corpus is the one I need?

Depending on which research question you have in mind using a corpus, it is important to consider the research questions the corpus builders had in mind constructing their corpus, i.e. whether you would find the phenomena you are looking for in that corpus. For example, you would hardly find forms of address in a corpus of Old English literature or in a corpus of Middle English medical texts. Rather, you would want to look at a corpus which contains letters, e.g. the Corpus of Early English Correspondence (CEEC) or at a corpus of spoken language. Therefore, it is always helpful to study the corpus descriptions first, in order to find out which types of texts the corpus contains and from which period they are.

3. How important/representative is the word count?

Depending on your research question, i.e. whether you plan on using a quantitative or a qualitative method, the word count could also be one important criterion to decide which corpus you should use. Using a qualitative method of analysis, a corpus with a low word count might be sufficient as one might want to meticulously look at each example and analyse it in its context. However, if it is statistical (hence quantitative information) evidence you are after, e.g. looking at how often a certain linguistic construction is used, it is important to use a corpus with a high word count and, if possible, a variety of text types, i.e. a reference corpus. This way you can make sure that the evidence you find is valid and not due to the style of a certain author (who by chance just likes to use that construction) or genre.

4. What does it mean when a corpus is annotated and how does that help me?

When a corpus is annotated it means that besides the actual words of the text, secondary information or metadata on the text is available. It also implies that you can search the corpus for this information. One example for annotation is lemmatisation. It means that each word in the corpus also contains information about its word stem, i.e. one can search the corpus for the word stems, the lemmas. Quite often corpora are annotated for their parts-of-speech (POS-tags), i.e. they contain information about the word classes of the words. This process is called part-of-speech tagging (POS-tagging) and the information on which categorisation of the word classes has been used and how the tags are abbreviated can be found checking the "tagset" in the corpus documentation. Furthermore, a corpus can contain information about the translation, glosses, phonetics, morphology, semantics, or pragmatics of its items; the latter of which are extremely helpful in corpora of spoken language, whilst translations and glosses are useful for the work on machine translation or teaching purposes. Quite often corpora have different layers of annotation in order to ease the use and accessibility of the annotated information. In which way the corpus has been annotated should also be stated in the corpus documentation. Finally, a corpus can be annotated for the syntax of its sentences. This annotation is called "parsing" and as the syntax is often parsed in a tree structure, these corpora are also called "treebanks".