DOWNLOAD PDF DEVELOPING LINGUISTIC CORPORA

Chapter 1 : Linguistic Corpora - Linguistics - Research Guides at UCLA Library

"where P(A) is the proportion of time that the coders agree and P(E) is the proportion of times that we would expect them to agree by chance." (Carletta 4).There is no doubt that annotation tends to be highly labour-intensive and time-consuming to carry out well.

Overview[edit] A corpus may contain texts in a single language monolingual corpus or text data in multiple languages multilingual corpus. Multilingual corpora that have been specially formatted for side-by-side comparison are called aligned parallel corpora. There are two main types of parallel corpora which contain texts in two languages. In a translation corpus, the texts in one language are translations of texts in the other language. In a comparable corpus, the texts are of the same kind and cover the same content, but they are not translations of each other. Machine translation algorithms for translating between two languages are often trained using parallel fragments comprising a first language corpus and a second language corpus which is an element-for-element translation of the first language corpus. Another example is indicating the lemma base form of each word. When the language of the corpus is not a working language of the researchers who use it, interlinear glossing is used to make the annotation bilingual. Some corpora have further structured levels of analysis applied. In particular, a number of smaller corpora may be fully parsed. Such corpora are usually called Treebanks or Parsed Corpora. The difficulty of ensuring that the entire corpus is completely and consistently annotated means that these corpora are usually smaller, containing around one to three million words. Other levels of linguistic structured analysis are possible, including annotations for morphology , semantics and pragmatics. Corpora are the main knowledge base in corpus linguistics. The analysis and processing of various types of corpora are also the subject of much work in computational linguistics , speech recognition and machine translation , where they are often used to create hidden Markov models for part of speech tagging and other purposes. Corpora and frequency lists derived from them are useful for language teaching. Corpora can be considered as a type of foreign language writing aid as the contextualised grammatical knowledge acquired by non-native language users through exposure to authentic texts in corpora allows learners to grasp the manner of sentence formation in the target language, enabling effective writing. Some archaeological corpora can be of such short duration that they provide a snapshot in time. One of the shortest corpora in time, may be the 15â€”30 year Amarna letters texts BC. Some notable text corpora[edit].

Chapter 2 : Corpus linguistics - Wikipedia

A linguistic corpus is a collection of texts which have been selected and brought together so that language can be studied on the computer. Today, corpus linguistics offers some of the.

What is corpus annotation? Corpus annotation is the practice of adding interpretative linguistic information to a corpus. For example, one common type of annotation is the addition of tags, or labels, indicating the word class to which words in a text belong. This is so-called part-of-speech tagging or POS tagging , and can be useful, for example, in distinguishing words which have the same spelling, but different meanings or pronunciation. The meanings of these same-looking words are very different, and also there is a difference of pronunciation, since the verb present has stress on the final syllable. Using one simple method of representing the POS tags â€" attaching tags to words by an underscore symbol â€" these three words may be annotated as follows: For others, annotation is a means to make a corpus much more useful â€" an enrichment of the original raw corpus. In this chapter, I will assume that such annotation is a benefit, so long as it is done well, with an eye to the standards that ought to apply to such work. Apart from part-of-speech POS tagging, there are other types of annotation, corresponding to different levels of linguistic analysis of a corpus or text â€" for example: For further information on such kinds of annotation, see Garside et al. In fact, it is possible to think up untold kinds of annotation that might be useful for specific kinds of research. One example is dysfluency annotation: Another illustration comes from an area of corpus research which has flourished in the last ten years: A glance at some of the advantages of an annotated corpus will help us to think about the standards of good practice these corpora require. Manual examination of a corpus What has been built into the corpus in the form of annotations can also be extracted from the corpus again, and used in various ways. For example, one of the main uses of POS tagging is to enhance the use of a corpus in making dictionaries. Thus lexicographers, searching through a corpus by means of a concordancer, will want to be able to distinguish separate verb from separate adjective , and if this distinction is already signalled in the corpus by tags, the separation can be automatic, without the painstaking search through hundreds or thousands of examples that might otherwise be necessary. Equally, a grammarian wanting to examine the use of progressive aspect in English is working, has been eating, etc can simply search, using appropriate search software, for sequences of BE any form of the lemma followed â€" allowing for certain possibilities of intervening words â€" by the ing-form of a verb. Automatic analysis of a corpus Similarly, if a corpus has been annotated in advance, this will help in many kinds of automatic processing or analysis. For example, corpora which have been POS-tagged can automatically yield frequency lists or frequency dictionaries with grammatical classification. Such listings will treat leaves verb and leaves noun as different words, to be listed and counted separately, as for most purposes they should be. Another important case is automatic parsing, i. Thirdly, consider the case of speech synthesis: Re-usability of annotations Some people may say that the annotation of a corpus for the above cases is not needed, automatic processing could include the analysis of such features as part of speech: This argument may work for some cases, but generally the annotation is far more useful if it is preserved for future use. The fact is that linguistic annotation cannot be done accurately and automatically: This is far from ideal, and to obtain an optimally tagged corpus, it is necessary to undertake manual work, often on a large scale. The automatically tagged corpus afterwards has to be post-edited by a team of human beings, who may spend thousands of hours on it. The result of such work, if it makes the corpus more useful, should be built into a tagged version of the corpus, which can then be made available to any people who want to use the tagging as a springboard for their own research. The BNC itself â€" all million words of it â€" has been automatically tagged but has not been manually post-edited, as the expense of undertaking this task would be prohibitive. In short, an annotated corpus is a sharable resource, an example of the electronic resources increasingly relied on for research and study in the humanities and social sciences. Multi-functionality If we take the re-usability argument one step further, we note that annotation often has many different purposes or applications: This has already been illustrated in the case of POS tagging: People who build corpora are familiar with the idea that no one in their right mind would offer to predict the future uses of a corpus â€"

future uses are always more variable than the originator of the corpus could have imagined! The same is true of an annotated corpus: However, this multi-functionality argument does not always score points for annotated corpora. There is a contrary argument that the annotations are more useful, the more they are designed to be specific to a particular application. Useful standards for corpus annotation What I have said above about the usefulness of annotated corpora, of course, depends crucially on whether the annotation has been well planned and well carried out. It is important, then, to recommend a set of standards of good practice to be observed by annotators wherever possible. It should always be easy to separate the annotations from the raw corpus, so that the raw corpus can be retrieved exactly in the form it had before the annotations were added. This is common sense: Detailed and explicit documentation should be provided Lou Burnard in chapter 3 emphasises the need to provide adequate documentation about the corpus and its constituent texts. For similar reasons, it is important to provide explicit and detailed documentation about the annotations in an annotated corpus. How, where, when and by whom were the annotations applied? Mention any computer tools used, and any phases of revision resulting in new releases, etc. What annotation scheme was applied? An annotation scheme is an explanatory system supplying information about the annotation practices followed, and the explicit interpretation, in terms of linguistic terminology and analysis, for the annotation. This is very important â€" Section 6 below will deal with annotation schemes. What coding scheme was used for the annotations? By coding scheme, I mean the set of symbolic conventions employed to represent the annotations themselves, as distinct from the original corpus. Again, I will devote a separate section to this Section 5. How good is the annotation? It might be thought that annotators will always proclaim the excellence of their annotations. Annotators should supply what information they can on the quality of the annotation. Arguably, the annotation practices should be linguistically consensual This and the following maxims are more open to debate. Any type of annotation presupposes a typology â€" a system of classification â€" for the phenomena being represented. But linguistics, like most academic disciplines, is sadly lacking in agreement about the categories to be used in such description. Different terminologies abound, and even the use of a single term, such as verb phrase, is notoriously a prey to competing theories. Even an apparently simple matter, such as defining word classes POS , is open to considerable disagreement. Against this background, it might be suggested that corpus annotation cannot be usefully attempted: However, looking at linguistics more carefully, we can usually observe a certain consensus: This is likely to be useful for other users and therefore to fit in with the re-usability goal for annotated corpora. Significantly, this consensual approach to categories is found not only in annotated corpora, but also in another key kind of linguistic resource â€" dictionaries. If, on the other hand, an annotator were to use categories specific to a particular theory and out of line with other theories, the annotated corpus would suffer in being less useful as a sharable resource. Annotation practices should respect emergent de facto standards This principle of good practice may be seen as complementary to the preceding one. By de facto standards, I mean some kind of standardisation that has already begun to take place, due to influential precedents or practical initiatives in the research community. De facto standards, on the other hand, emerge often gradually from the research community in a bottom-up manner. De facto standards encapsulate what people have found to work in the past, which argues that they should be adopted by people undertaking a new research project, to support a growing consensus in the community. However, often a new project breaks new ground, for example with a different kind of data, a different language, a different purpose those of previous projects. It would clearly be a recipe for stagnation if we were to coerce new projects into the following exactly the practices of earlier ones. Nevertheless it makes sense for new projects to respect the outcomes of earlier projects, and only to depart from their practices where this can be justified. In 8 below, I will refer to some of the incipient standards for different kinds of annotation and mark-up. These can only be presented tentatively, however, as the practice of corpus annotation is continually evolving. In the early s, the European Union launched an initiative under the name of EAGLES Expert Advisory Groups on Language Engineering Standards with the goal of encouraging standardisation of practices for natural language processing in academia and industry, particularly but not exclusively in the EU. The encoding of annotations But before focussing on annotation schemes and the linguistic categories they incorporate, it will be helpful to touch briefly on the encoding of annotations â€" that is, the actual symbolic representations used. This means

we are for the moment concentrating on how annotations are outwardly manifested â€" for example, what you see when you inspect a corpus file on your computer screen â€" rather than what their meaning is, in linguistic terms. The presentation of the tag itself may be complex or simple. One basic requirement is that the POS tag or any other annotation device should be unambiguous in representing what it stands for. Another requirement, useful for everyday purposes such as reading a concordance on a screen, is brevity: A third requirement, more useful in some contexts than in others, is that the annotation device should be transparent to the human reader rather than opaque. The example NP1 is at least to some degree intelligible, and is less mystifying than it would be if some arbitrary sequence of symbols, say Q! The type of tag illustrated above originated with the earliest corpus to be POS-tagged in , the Brown Corpus. More recently, since the early s, there has been a far-reaching trend to standardize the representation of all phenomena of a corpus, including annotations, by the use of a standard mark-up language â€" normally one of the series of related languages SGML, HTML, and XML see Lou Burnard, chapter 3. One advantage of using these languages for encoding features in a text is that they provide a general means of interchange of documents, including corpora, between one user or research site and another. Furthermore, the use of the mark-up language itself can be efficiently parsed or validated, enabling the annotator to check whether there are any ill-formed traits in the markup, which would signal errors or omissions. Yet another advantage is that, as time progresses, tools of various kinds are being developed to facilitate the processing of texts encoded in these languages. This, however, would require a further step of processing which may not be easy to manage for the technically less adept user. Attempts have been made to make this type of logical encoding more accessible, by relaxing standards of conformance. Within the overall framework SGML, different co-existing encoding standards have been proposed or implemented: Any corpus of spoken data, in particular, is likely to contain such cross-bracketing, for example in the cross-cutting of stretches of speech which need to be marked for different levels of linguistic information â€" such phenomena as non-fluencies, interruptions, turn overlaps, and grammatical structure are prone to cut across one another in complex ways. But because of the difficulties I have mentioned, many people will find it easier meanwhile to follow the lead of other well-known encoding schemes â€" such as the simpler styles of mark-up associated with the Brown and ICE families of corpora, or with the CHILDES database of child language data. As the name suggests, CHILDES is neither a corpus nor a coding scheme in itself, but it provides both, operating as a service which pools together the data of many researchers all over the world, using a common coding and annotation schemes, and common software including annotation software. Annotation manual Why do we need an annotation manual? This document is needed to explain the annotation scheme to the users of an annotated corpus. Typically such manuals originate from sets of guidelines which evolve in the process of annotating a corpus â€" especially if hand editing of the corpus has been undertaken. A most carefully worked-out annotation scheme was published as a weighty book by Geoffrey Sampson Although annotation manuals often build up piecemeal in this way, for the present purpose we should see them as completed documents intended for corpus users. They can be thought of as consisting of two sections â€" a a list of annotation devices and b a specification of annotation practices â€" which I will illustrate, as before, using the familiar case of a POS tagging scheme for an example, see Johansson, , for the LOB Corpus, or Sampson, , Ch. A list of annotation devices with brief explanations This list acts as a glossary â€" a convenient first port of call for people trying to make sense of the annotations.

Chapter 3 : Developing linguistic corpora : a guide to good practice (Book,) [calendrierdelascience.com]

A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research.

Similarly, the early Arabic grammarians paid particular attention to the language of the Quran. In the Western European tradition, scholars prepared concordances to allow detailed study of the language of the Bible and other canonical texts. Nelson Francis of Computational Analysis of Present-Day American English in , a work based on the analysis of the Brown Corpus , a carefully compiled selection of current American English, totalling about a million words drawn from a wide variety of sources. The AHD took the innovative step of combining prescriptive elements how language should be used with descriptive information how it actually is used. Other publishers followed suit. Other corpora represent many languages, varieties and modes, and include the International Corpus of English , and the British National Corpus , a million word collection of a range of spoken and written texts, created in the s by a consortium of publishers, universities Oxford and Lancaster and the British Library. An example is the Andersen -Forbes database of the Hebrew Bible, developed since the s, in which every clause is parsed using graphs representing up to seven levels of syntax, and every segment tagged with seven fields of information. This is a recent project with multiple layers of annotation including morphological segmentation, part-of-speech tagging , and syntactic analysis using dependency grammar. Methods[edit] Corpus linguistics has generated a number of research methods, which attempt to trace a path from data to theory. Wallis and Nelson [10] first introduced what they called the 3A perspective: Annotation, Abstraction and Analysis. Annotation consists of the application of a scheme to texts. Annotations may include structural markup, part-of-speech tagging, parsing, and numerous other representations. Abstraction consists of the translation mapping of terms in the scheme to terms in a theoretically motivated model or dataset. Abstraction typically includes linguist-directed search but may include e. Analysis consists of statistically probing, manipulating and generalising from the dataset. Analysis might include statistical evaluations, optimisation of rule-bases or knowledge discovery methods. Most lexical corpora today are part-of-speech-tagged POS-tagged. In such situations annotation and abstraction are combined in a lexical search. The advantage of publishing an annotated corpus is that other users can then perform experiments on the corpus through corpus managers. By sharing data, corpus linguists are able to treat the corpus as a locus of linguistic debate, rather than as an exhaustive fount of knowledge. Recent studies have suggested treatment outcome in adolescents with social anxiety disorder can also be assessed by analysing language by means of Corpus Linguistics.

Chapter 4 : Corpus of Contemporary American English (COCA)

In this volume, a selection of leading experts in various key areas of corpus construction offer advice in a readable and largely non-technical style to help the reader to ensure that their corpus is well designed and fit for the intended purpose.

The guiding principles that relate corpus and text are concepts that are not strictly definable, but rely heavily on the good sense and clear thinking of the people involved, and feedback from a consensus of users. However unsteady is the notion of representativeness, it is an unavoidable one in corpus design, and others such as sample and balance need to be faced as well. It is probably time for linguists to be less squeamish about matters which most scientists take completely for granted. I propose to defer offering a definition of a corpus until after these issues have been aired, so that the definition, when it comes, rests on as stable foundations as possible. For this reason, the definition of a corpus will come at the end of this paper, rather than at the beginning. Who builds a corpus? Experts in corpus analysis are not necessarily good at building the corpora they analyse â€" in fact there is a danger of a vicious circle arising if they construct a corpus to reflect what they already know or can guess about its linguistic detail. Ideally a corpus should be designed and built by an expert in the communicative patterns of the communities who use the language that the corpus will mirror. Quite regardless of what is inside the documents and speech events, they should be selected as the sorts of documents that people are writing and reading, and the sorts of conversations they are having. Factual evidence such as audience size or circulation size can refine such sampling. The corpus analyst then accepts whatever is selected. This could be stated as a principle: 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. Obviously if it is already known that certain text types contain large numbers of a microlinguistic feature such as proper nouns or passive verb phrases, it becomes a futile activity to "discover" this by assembling a corpus of such texts. Selection criteria that are derived from an examination of the communicative function of a text are called external criteria, and those that reflect details of the language of the text are called internal criteria. Corpora should be designed and constructed exclusively on external criteria Clear 1. What is a corpus for? A corpus is made for the study of language; other collections of language are made for other purposes. So a well-designed corpus will reflect this purpose. The contents of the corpus should be chosen to support the purpose, and therefore in some sense represent the language from which they are chosen. Since electronic corpora became possible, linguists have been overburdened by truisms about the relation between a corpus and a language, arguments which are as irrelevant as they are undeniably correct. Everyone seems to accept that no limits can be placed on a natural language, as to the size of its vocabulary, the range of its meaningful structures, the variety of its realisations and the evolutionary processes within it and outside it that cause it to develop continuously. Therefore no corpus, no matter how large, how carefully designed, can have exactly the same characteristics as the language itself. So we sample, like all the other scholars who study unlimitable phenomena. We remain, as they do, aware that the corpus may not capture all the patterns of the language, nor represent them in precisely the correct proportions. In fact there are no such things as "correct proportions" of components of an unlimited population. Corpus builders should strive to make their corpus as representative as possible of the language from which it is chosen. However hard we strive, a corpus will occasionally show features which we suspect not to be characteristic of the language under study, or fail to show features which are expected. Following our first principle above, we should not feel under pressure to use the patterns of the language to influence the design of the corpus, but we should review the design criteria to check that they are adequate. To optimise the application of this principle we can make use of an important resource within ourselves, which is not available to most scientific researchers in other disciplines. As sophisticated users of at least one language, we have an inbuilt awareness of language structure, often called intuition, that gives a personal, independent and non-negotiable assessment of language pattern. Intuition can help in many ways in language research, in conjunction with other criteria of a more examinable nature. The drawbacks to intuition are a that we cannot justify its use beyond personal testimony, and b that people differ notoriously in their intuitive judgements. In this context we should also be aware that

an incautious use of intuition in the selection of texts for a corpus would undermine the first principle 2. There are three considerations that we must attend to in deciding a sampling policy: The orientation to the language or variety to be sampled. The criteria on which we will choose samples. The nature and dimensions of the samples. Orientation This is not a crisply delineated topic, and has largely been taken for granted so far in corpus building. The early corpora, for example the Brown corpus and those made on its model Hofland and Johansson , were normative in their aims, in that their designers wanted to find out about something close to a standard language. The word "standard" appears in the original Brown title; by choosing published work only, they automatically deselected most marked varieties. Most of the large reference corpora of more recent times adopt a similar policy; they are all constructed so that the different components are like facets of a central, unified whole. Such corpora avoid extremes of variation as far as possible, so that most of the examples of usage that can be taken from them can be used as models for other users. Some corpora have a major variable already as part of the design â€" a historical corpus, for example, is deliberately constructed to be internally contrastive, not to present a unified picture of the language over time though that could be an interesting project. Another kind of corpus that incorporates a time dimension is the monitor corpus Sinclair ; a monitor corpus gathers the same kind of language at regular intervals and its software records changes of vocabulary and phraseology. Parallel corpora, or any involving more than one language, are of the same kind â€" with inbuilt contrasting components; so also is the small corpus used in Biber et. These corpora could be called contrastive corpora because the essential motivation for building them is to contrast the principal components. There is a guiding principle here of great importance, and one which is commonly ignored. Only those components of corpora which have been designed to be independently contrastive should be contrasted. That is to say, the existence of components differentiated according to the criteria discussed below, or identified by archival information, does not confer representative status on them, and so it is unsafe to use them in contrast with other components. Now that with many corpus management systems it is possible to "dial-a-corpus" to your own requirements, it is important to note that the burden of demonstrating representativeness lies with the user of such selections and not with the original corpus builder. It is perfectly possible, and indeed very likely, that a corpus component can be adequate for representing its variety within a large normative corpus, but inadequate to represent its variety when freestanding. This point cannot be overstated; a lot of research claims authenticity by using selections from corpora of recognised standing, such as the Helsinki Corpus, which is a notable reference corpus covering the language of almost a millennium in a mere 1,, words. Each small individual component of such a corpus makes its contribution to the whole and its contrasts with other segments, but was never intended to be a freestanding representative of a particular state of the language. See the detailed description at http: Normative, historical, monitor and varietal corpora are not the only kinds; demographic sampling has been used a little, and there are all sorts of specialised corpora. For an outline typology of corpus and text see Sinclair , which is a summary and an update of a report made for the European Commission for that report see the EAGLES server at http: Criteria Any selection must be made on some criteria and the first major step in corpus building is the determination of the criteria on which the texts that form the corpus will be selected. Often some of these large-scale criteria are pre-determined by constraints on the corpus design â€" for example a corpus called MICASE stands for the Michigan Corpus of Academic Spoken English, and the corpus consists of speech events recorded on the Ann Arbor campus of the University of Michigan on either side of the millennium; it follows that the language in the corpus will mainly be of the large variety called American English. All the above criteria are pre-determined, and all but the date are built into the name of this corpus, so its own structural criteria will be set at a more detailed level 3. All but the most comprehensive corpora are likely to use one or more criteria which are specific to the kind of language that is being gathered, and it is not possible to anticipate what these are going to be. The corpus designer should choose criteria that are easy to establish, to avoid a lot of labour at the selection stage, and they should be of a fairly simple kind, so that the margin of error is likely to be small. If they are difficult to establish, complex or overlapping they should be rejected, because errors in classification can invalidate even large research projects and important findings. Now that there are a number of corpora of all kinds available, it is helpful to look at the criteria that have been used, and to evaluate them in three ways â€" as themselves, how

useful and valuable a variety of the language they depict; as a set of criteria, how they interact with each other and avoid ambiguity and overlap; and the results that they give when applied to the corpus. 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. Beyond these criteria it is possible to envisage an unlimited categorisation of people, places and events, any of which are potentially valuable for one study or another see the typology mentioned above. The gender of the originator of a text has been a popular criterion in recent years, though few texts have a single originator whose gender is known, and hoaxes are not unknown for example it was recently alleged that the works of a famous crime writer noted for rough-and-tough stories were in fact composed by his wife. It is essential in practice to distinguish structural criteria from useful information about a text. For a corpus to be trusted, the structural criteria must be chosen with care, because the concerns of balance and representativeness depend on these choices. Other information about a text can, of course, be stored for future reference, and scholars can make up their own collections of texts to suit the objectives of their study. The question arises as to how and where this information should be stored, and how it should be made available. Because it is quite commonly added to the texts themselves, it is an issue of good practice, especially since in some cases the additions can be much larger than the original texts. In the early days of archiving text material, the limitations of the computers and their software required a structurally simple model; also before there was an abundance of language in electronic form, and before the internet made it possible for corpora to be accessed remotely, it was necessary to agree protocols and practices so that data could be made available to the research community. The model that gained widest acceptance was one where additional material was interspersed in the running text, but enclosed in diamond brackets so that it could â€” at least in theory â€” be found quickly, and ignored if the text was required without the additions. Nowadays there is no need to maintain a single data stream; modern computers have no difficulty storing the plain text without any additions, and relating it token by token to any other information set that is available, whether "mark-up", which is information about the provenance, typography and layout of a printed document, or "annotation", which is analytic information usually about the language 4. It is also possible nowadays to store facsimiles of documents and digitised recordings of speech, and have the computer link these, item by item, to plain text, thus removing even the need to have mark-up at all. 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. Sampling Looking down from the totality of the corpus, the major criteria will define several components, while at the other end are the individual texts, which will be such things as written or printed documents, and transcripts of spoken events. Cells are the groupings formed from the intersection of criteria. The first-level components will be small in number, for practical reasons, because if there are too many then either each component will be very small or the corpus will be very large. The simplest classification is binary, so that if a corpus of spoken language is first divided into "private" and "public", then each of these types will have to be represented by a sufficiently large amount of text for its characteristics to become evident. If the next criterion is "three or fewer active participants", as against "more than three active participants", then each of the original categories is divided into two, and the theoretical size of the corpus doubles. Each criterion divides the corpus into smaller cells; if we assume that the criteria are binary and cross-cutting then as we have just seen two criteria divide the corpus into four cells, three into eight, four into sixteen etc. You then have to decide what is the acceptable minimum number of words in a cell; this depends quite a lot on the type of study you are setting out to do, but if it is not substantial then it will not supply enough reliable evidence as part of the overall picture that the corpus gives of the language. This is known as the "scarce data problem". The matter of size is discussed later, and the example in the following paragraph is only illustrative. If you decide on, say, a million words as the minimum for a cell, then with four criteria you need a corpus with a minimum size of sixteen million words. Each additional binary criterion doubles the minimum size of the corpus, and in addition we find that real life is rarely as tidy as this model suggests; a corpus where the smallest cell contains a million words is likely in practice to have several cells which contain much more. This involves the question of balance, to which we will return. There are also questions of criteria that have more than two options, and

of what to do with empty or underfilled cells, all of which complicate the picture. The matter of balance returns as we approach the smallest item in a corpus, the text. Here arises another issue in sampling that affects, and is affected by, the overall size of the corpus. Language artefacts differ enormously in size, from a few words to millions, and ideally, documents and transcripts of verbal encounters should be included in their entirety. The problem is that long texts in a small corpus could exert an undue influence on the results of queries, and yet it is not good practice to select only part of a complete artefact. However it is an unsafe assumption that any part of a document or conversation is representative of the whole â€" the result of research for decades of discourse and text analysis make it plain that position in a communicative event affects the local choices. The best answer to this dilemma is to build a large enough corpus to dilute even the longest texts in it. If this is not practical, and there is a risk that a single long text would have too great an influence on the whole, so recourse has to be made to selecting only a part of it, and this has to be done on "best guess" grounds. But even a very large corpus may find it almost impossible to get round copyright problems if the builders insist on only complete texts. The rights holders of a valuable document may not agree to donate the full text to a corpus, but if it is agreed that occasional passages are omitted, so that the value of the document is seriously diminished, then the rights holders might be persuaded to relent.

Chapter 5 : Developing Linguistic Corpora

A linguistic corpus is a collection of texts which have been selected and brought together so that language can be studied on the computer. Today, corpus linguistics offers some of the most powerful new procedures for the analysis of language, and the impact of this dynamic and expanding sub-discipline is making itself felt in many areas of language study.

Chapter 6 : Linguistic Corpora - Linguistics - Library Guides at UChicago

Developing linguistic theories using annotated corpora 3 Intuition and experiment Experimental and corpus methods are often deï¬•ned in opposition to 'intuition-.

Chapter 7 : LINGUIST List Developing Linguistic Corpora, Available Online

Usage-based linguistic studies have gained new insights as corpus-based and corpus-driven analyses have advanced in recent years. Linguists working in different domains have turned to corpora as a.

Chapter 8 : Text corpus - Wikipedia

In recent years, corpus-based research has gained considerable momentum in linguistics, overtaking arm chair type observations on natural language with a more e.

Chapter 9 : BYU corpora: billions of words of data: free online access

Linguistic Corpora: A collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting-point of linguistic description or as a means of verifying hypotheses about a language (corpus linguistics).