Linguistic annotation in/for corpus linguistics

Stefan Th. Gries & Andrea L. Berez
University of California, Santa Barbara and University of Hawai'i at Mānoa

Abstract
This article surveys linguistic annotation in corpora and corpus linguistics. We first define the concept of 'corpus' as a radial category and then, in Section 2, discuss a variety of kinds of information for which corpora are annotated and that are exploited in contemporary corpus linguistics. Section 3 then exemplifies many current formats of annotation with an eye to highlighting both the diversity of formats currently available and the emergence of XML annotation as, for now, the most widespread form of annotation. Section 4 summarizes and concludes with desiderata for future developments.

1 Introduction

1.1 Definition of a corpus

This chapter is concerned with the use of linguistic annotation for corpus-linguistic analyses. It is therefore useful to begin with a brief definition of the notion of corpus, especially since scholars differ in how freely or conservatively they apply this notion. We consider the notion of *corpus* to constitute a radial category of the same kind as a polysemous word. That is, it is a category that contains exemplars that are prototypical by virtue of exhibiting several widely accepted characteristics, but that also contains many exemplars that are related to the prototype or, less directly, to other exemplars of the category by family resemblance links.

The characteristics that jointly define a prototypical corpus are the following: the corpus

- consists of one or more *machine-readable* Unicode text files (although, even as late as in Tagliamonte (2007:226), one still finds reference to corpora as ASCII files);[1]
- is meant to be *representative* for a particular kind of speaker, register, variety, or language as a whole, which means that the sampling scheme of the corpus represents the variability of the population it is meant to represent;
- is meant to be *balanced*, which means that the sizes of the subsamples (of speakers, registers, varieties) are proportional to the proportions of such speakers, registers, varieties, etc. in the population the corpus is meant to represent; and
- contains data from *natural communicative settings*, which means that at the time the language data in the corpus were produced, they were not produced solely for the purpose of being entered into a corpus, and/or that the production of the language data was as untainted by the collection of those data as possible.

Given these criteria, it is probably fair to say that the British National Corpus (BNC) represents a prototypical corpus: its most widely used version, the BNC World Edition XML, consists of 4049 XML-annotated Unicode text files (containing altogether approximately 100m words) that are intended to be representative of British English of the 1990s. Furthermore, these files contain one of the largest sections of spoken data available (10m words), to be

1 A reviewer points out that most corpora are in English and are thus by default Unicode-compliant, since English orthographic characters use the ASCII subset of Unicode.

representative of the importance of spoken language in our daily lives.

Less prototypical corpora differ from the prototype along one or more of the above main criteria, or along other, less frequent criteria. For example, many new corpora are not just based on texts, but on audio and/or video recordings, which gives rise to many challenges regarding transcription and annotation (see below). However, the greatest variation between corpora probably regards the criterion of natural communicative setting, which gives rise to many different degrees of naturalness and, thus, results in different corpora occupying different places in the multidimensional space of experimental and observational data (cf. Gries 2013 for a three-dimensional model space of linguistic data). For example, the following corpora involve slightly less natural settings:

- the Switchboard Corpus (Godfrey & Holliman 1997) contains telephone conversations between strangers on assigned topics – while talking on the phone is a normal aspect of using language, talking to strangers about assigned topics is not.
- the International Corpus of Learner English (Granger et al. 2002) contains timed and untimed essays written by foreign language learners of English on assigned topics – while writing about a topic is a fairly normal aspect of using language, writing on an assigned topic under time pressure is not (outside of instructional settings).

In some sense, corpora consisting of newspaper texts and web data are even less prototypical corpora. While such corpora are often vast and relatively easy to compile, they can represent quite particular registers: for instance, newspaper articles are created more deliberately and consciously than many other texts, they often come with linguistically arbitrary restrictions regarding, say, word or character lengths, they are often not written by a single person, they may be heavily edited by editors and typesetters for reasons that again may or may not be linguistically motivated, etc. Many of these conditions may also apply to (some) web-based corpora, although web corpora are increasingly becoming more frequent examples of written language use.

Other corpora are documentary-linguistic in nature, designed to provide an overview of an understudied, small, or endangered language before the language ceases to be spoken. These corpora are usually considerably smaller than the prototypical corpus and are based on audio and video recordings that are transcribed, annotated, and described with metadata by either a single researcher working in the field or by a small team of researchers (Himmelmann 2006 terms the recordings the *primary data* of a documentary corpus, while the transcription, annotation, and descriptive metadata are known as the *apparatus* of the corpus). The theorization of documentary linguistic corpora is often less straightforward than that of a prototypical corpus, since it may be difficult to get a balanced or representative corpus of a language undergoing community-wide attrition; in addition, the stakeholders in the corpus may be a relatively small group of academic linguists and/or language community members, and local politics and culturally-determined ethical obligations will likely play a role in the ultimate contents of a documentary corpus (see, e.g. Czaykowska-Higgins 2009, Woodbury 2011, Rice 2012). Nonetheless, corpus linguistic and documentary methods of annotation overlap in both practice and motivation, and are thus included here.

Finally, there are corpora that are decidedly experimental in nature, and thus 'violate' the criterion of natural communicative setting even more. An extreme example, Bard et al. (1996), compiled the DCIEM Map Task Corpus, which consists of task-oriented, unscripted dialogs in which one interlocutor describes a map route to the other, after both interlocutors had been subjected to 60 hours of sleep deprivation and to one of three drug treatments. Another example

is the TIMIT corpus (Garofolo et al. 1993), which contains broadband recordings of 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences.

1.2 What do corpus linguists do with corpora?

Given the above-mentioned diversity and task-specificity of corpora, it should come as no surprise that many different annotation types and formats are used in corpus linguistics. In spite of the large number of different uses, much of corpus linguistics is still dominated by a relatively small number of application types – in spite of calls to arms by, say, McEnery & Ostler (2000), it is only in the last few years that more and more corpora are compiled and annotated for non-English data and for more than the 'usual' high-frequency applications. According to a survey by Gilquin & Gries (2009), corpus-linguistic studies published over the course of four years in three major corpus-linguistic journals were mostly

- exploratory (as opposed to hypothesis-testing) in nature;
- on matters of lexis and phraseology, followed by syntax;
- based on written data;
- using frequency data and concordances, followed by simple association measures.

Given the predominance of such applications, it comes as no surprise that the most commonly found kind of annotation is part-of-speech tagging. However, over the last 20 years, many corpora have begun to feature other kinds of annotation. In the next section, we provide a survey of the kinds of information that corpora may be annotated for. In this survey, we are less concerned with markup in the sense that it is often used in corpus linguistics to denote metadata about a corpus file, which might include information like when the data were collected, a description of the data source, when the file was prepared, demographic information about participants, and the like. Rather, we will focus on markup as annotation proper, i.e. information/elements added to provide specifically linguistic/grammatical/structural information such as part of speech, semantics, pragmatics, prosody, interaction and many others.

2 What are corpora annotated for?

The types of information corpora are annotated for is dependent on the kind, and thus typicality, of corpus, i.e. the way in which the data have been collected. Obviously, just about every corpus can be annotated for part-of-speech and/or lemma information, whereas many corpora do not easily allow for other kinds of annotation. For example, many written corpus data in general can be annotated for the identity of the author but cannot be annotated for prosodic, gestural, or interactional aspects of language production. By contrast, conversations between speakers that are video-taped and transcribed can be annotated for a large variety of linguistic and contextual information, although usually not all the information that an audio/video recording contains can be unambiguously annotated, given how costly annotation often is in terms of time and resources, and how widely research questions, objectives, and strategies differ from one researcher to the next, and from one project to the next. In this section, we provide an overview of linguistic and paralinguistic information that corpus linguists frequently use in their work.

2.1 Frequent forms of annotation of written corpora

In this section, we are concerned with annotation that describes inherently linguistic

characteristics of the language sample in the corpus. This kind of annotation requires an initial segmentation process called *tokenization*, which aims to determine and delineate the units in the corpus that will be annotated – words, numbers, punctuation marks, etc. In some cases, this involves an additional step called *named entity recognition*, which serves to determine the units in the corpus that are proper names. We will not discuss these here in more detail; cf. Schmid (2008) for discussion about multiwords in general.

2.1.1 Lemmas

One of the most basic types of annotation is *lemmatization*, the process of identifying and marking each word in a corpus with its base (citation or dictionary) form. In an English corpus this would involve, for example, stripping away inflectional morphology on verbs so that all forms of the lemma FORGET – *forget*, *forgets*, *forgetting*, *forgot*, and *forgotten* – would be marked as representing a form of FORGET, and could be retrieved without the user having to enter all forms of FORGET individually. Lemmatization can be performed on the basis of an existing form-lemma database, a (semi-)automatic approach called *stemming* in which word forms are truncated by cutting off characters to arrive at the more general representation of a lemma, or some hybrid approaches of these two strategies that may also involve morphological and/or syntactic analysis to disambiguate ambiguous forms (cf. Fitschen & Gupta 2008 for discussion).

2.1.2 Part-of-speech tagging: syntactic and morphological annotation

Part-of-speech tagging is one of the most frequent and most exploited kinds of annotation because it is relevant to many corpus-linguistic studies and because it feeds into many other annotation processes like lemmatization, syntactic parsing, semantic annotation etc. It involves assigning to each tokenized word a label that minimally identifies the part of speech of the word but that typically also includes some grammatical category information. For example, part-of-speech tags in English corpora often not only annotate the word *run* in *I regularly run marathons* as a verb, but also as a verb in the base form, thus distinguishing it from the infinitival *run* in *I am going to run a marathon*; many relatively standardized annotation formats for part-of-speech tags are available and are discussed below.

The precision of automatic part-of-speech annotation is highly dependent on many factors, including the language represented by the corpus and its morphological characteristics, the complexity of the text(s) in the corpus, the kind of tagger used (symbolic or, more commonly now, statistical), the size and precision of the corpora the tagger has been trained on, the size of the tagset, etc. As Charniak (1997:4) points out, however, for English one may already achieve a precision of approximately 90% just by assigning (i) to every word attested in the training corpus its most frequent part-of-speech tag and (ii) to every word attested that is not in the training corpus the tag *proper noun*. More sophisticated taggers for English corpora by now achieve precision in excess of 95% (cf. Schmid 2008:547), but tagging still runs into many problems in both morphologically relatively impoverished languages like English and in languages with relatively rich morphology. As for the former, some uses of words may genuinely be ambiguous (a famous example from the tagging guidelines of the Penn Treebank is the categorial status of *entertaining* in *The Duchess was entertaining last night*; cf. Santorini 1990:32). As for the latter, in morphologically richer languages, including morphological information in part-of-speech tags quickly inflates the inventory of required tags to such a degree that, for heavily polysynthetic languages, it may be impossible to devise and then apply an inventory of part-of-speech tags with any reasonable degree of precision. For example, it seems hard to imagine a tagset that can usefully deal with languages such as Dena'ina (Athabaskan) which has up to 19 prefix positions

before the verb stem – a tagset that can tag all the possible combinations of how these slots are filled is certainly conceivable but also likely to be unwieldy.

2.1.3 Syntactic parse trees

The annotation of corpora for *syntactic analyses with parse trees* followed part-of-speech tagging. The first corpora featuring parse trees were the Gothenburg Corpus, the SUSANNE Corpus, and the Lancaster Parsed Corpus (Zinsmeister et al. 2008:760), which involved either completely manual annotation, or the manual checking of the results of automatic parsing. Over the last decades, just like POS-tagging, syntactic parsing has evolved from symbolic approaches to statistical approaches that assign the most probable syntactic analyses, where the probability of a syntactic analysis is determined on the basis of a training corpus (supervised training) or an entirely data-driven process (unsupervised training). The results of such analyses come in the form of either phrase-structure representations – the most frequent parse type – or dependency-tree representations; often, the automatic analyses are post-processed manually to correct mistakes emerging from the automatic analysis.

A widely used example of a phrase-structure parsed corpus is the British Component of the International Corpus of English (ICE-GB; cf. Nelson, Wallis, & Aarts 2002), a one-million word corpus (60% spoken, 40% written data) representative for British English of the 1990s.This corpus is fully tagged for part-of-speech, syntactically parsed, and manually checked. Another well-known parsed corpus is the Penn Treebank (Marcus, Santorini, & Marcinkiewicz 1993) that contains materials from the Wall Street Journal corpus, the Switchboard corpus, and the Brown corpus and is currently available (from the Linguistic Data Consortium) in three differently annotated versions.

An example of a less widely-used but still very well-known parsed corpus is the TiGer corpus (Brants et al. 2004), of which the current version contains approximately 900K words / 50K sentences of German newspaper text. TiGer is freely available as plain text and in XML format with phrase-structure and dependency-structure representations.

In contemporary corpus-based research, the number of studies that rely on syntactically parsed corpora is steadily increasing. Given the higher error rates of fully automatic syntactic parsers as compared to part-of-speech taggers – even leaving aside the question of how parses by different parsers can be compared – however, many studies still involve large amounts of manual disambiguation and error checking. For example, researchers often query the syntactically parsed annotation of a corpus, but then still check each retrieved match (or a sizable sample of all matches) to ensure it really instantiates the intended syntactic structure. While this can be labor-intensive and may miss structures that the parser did not recognize/annotate as intended, it may still yield reasonable degrees of precision and recall. An alternative strategy that is also still widespread involves not utilizing the parse tree, but approximating the relevant syntactic construction by lexical and/or part-of-speech annotation only, which may result in perfect recall but which also requires a much larger number of matches to be checked for false hits. The two approaches can be contrasted on the basis of the so-called *into*-causative construction exemplified in (1).

(1) a. He [$_{VP}$ tricked [$_{NP\,DO}$ her] into [$_{VP}$ selling his car]].
b. She [$_{VP}$ bullied [$_{NP\,DO}$ him] into letting her [$_{VP}$ stay overnight]].

The former approach might aim at retrieving such examples on the basis of a parse tree query that describes the above structure of the VP (maybe including *into* in the description); the latter approach would involve retrieving all instance of *into* followed by a word (or verb, if part-

of-speech tags are available and used) ending in *ing*; the results of both queries would then be checked to identify true hits.

2.1.4 Semantic annotation

One frequent kind of semantic annotation relatively common in corpus linguistic studies involves the identification of senses of word forms in a corpus, which is often referred to as *word sense disambiguation*. Word sense disambiguation is often largely automatic and consists of an algorithm assigning to each word form a sense from an inventory of possible senses that best matches the context in which the word form is used. According to Rayson & Stevenson (2008), such algorithms are AI-based, knowledge-based, corpus-based, or a hybrid approach combining different techniques. However, the amount of published corpus-linguistic research that relies on automatic sense tagging appears to be quite small.

Another much less frequent scenario arises when researchers and their teams semantically annotate semantic phenomena like metaphor (or metonymy, synecdoche, etc.) in corpora. One well-known project to identify instances of metaphor in corpora is the Pragglejaz project headed by G. Steen, which resulted in a detailed annotation protocol called the Metaphor Identification Procedure that was applied to, for instance, the BNC Baby, a 4-million word sample from the British National Corpus.

Other projects that involve making available semantically-annotated corpus resources include the SenSem Corpus: an annotated corpus for Spanish and Catalan constructions with information about aspect, modality, polarity and factuality (<http://grial.uab.es/sensem/corpus/main>) or the TimeBank Corpus by Pustejovsky et al. (2003) containing "texts from various sources […] annotated with event classes, temporal information, and aspectual information" (Zinsmeister et al. 2008:762)

On many occasions, however, semantic annotation is done by individual researchers or teams for individual research projects. Such studies often involve non-standardized forms of annotation of a data set, and the resulting annotated data are often not shared with others. For example, in an attempt to explore the polysemy of the verb lemma RUN in corpus data, Gries (2006) studied more than 800 examples of RUN from two corpora to develop a network of senses. The analysis was based both on earlier cognitive-linguistic polysemy studies of (mostly) prepositions and a few other verbs and lexicographic resources such as corpus-informed dictionaries as well as the WordNet semantic database (Fellbaum 1998), which lists 41 different senses of the verb RUN.

While WordNet is one of the most widely-used semantic resources in corpus linguistics (though not a corpus itself), others are available including PropBank, FrameNet, and the UCREL Semantic Analysis System USAS. PropBank (Palmer, Gildea, & Kingsbury 2005) consists of "a layer of predicate-argument information, or semantic role labels, [that has been added] to the syntactic structures of the Penn Treebank" (p. 71) such that, for instance, roles such as agent, patient, etc. are distinguished verb-specifically.

FrameNet (<https://framenet.icsi.berkeley.edu/fndrupal/home>) is also not so much a corpus as a lexical corpus-based database containing more than 170K English sentences annotated for semantic roles of words as recognized in the theory of Frame Semantics (Fillmore 1976). While the database contains English data only, because frames are semantic in nature the resource is potentially also useful to researchers working on other languages. So far, FrameNet databases have been developed for Brazilian Portuguese, Chinese, German, Japanese, Spanish, and Swedish.

Finally, USAS is a semantic-analysis system that tags words in corpora as belonging to one of 21 semantic categories (e.g., general and abstract terms, the body and the individual,

linguistic actions, social actions, etc.) as well as additional more fine-grained subcategories (cf. Archer, Wilson, & Rayson 1992).

In spite of the importance and usefulness of semantic annotation for many areas of (corpus-)linguistic research – machine translation, information retrieval, content analysis, speech processing, discourse-pragmatic research on irony, corpus-based approaches to lexicography, etc. – it needs to be borne in mind that semantic annotation is an extremely time- and resource-consuming task. While humans seem to experience very little difficulty in accessing and understanding an appropriate sense of a word in natural communicative settings well enough for communication not to break down – both literal or metaphorical/idiomatic – humans tasked with *annotating* senses of words in context agree with each other less often than might be expected (cf. Fellbaum et al. 1998), as anyone who has ever tried to annotate senses of a word will confirm. Other reasons for, or correlates of, the difficulty of semantic annotation are that (i) it is not even clear whether there is really any such thing as discrete word senses (cf. Kilgarriff 1997) or whether uses of a word embody fuzzy meaning potentials that, while often effortlessly processable by humans, do not lend themselves to specific discretizing annotations; and that (ii) it is far from clear and/or specific to a particular project which level of resolution or granularity is most useful, since even dictionary senses differ considerably from the senses that linguistically naïve human subjects distinguish (Jorgensen 1990).

2.2 Forms of annotation of spoken/multimodal corpora

While most available corpora contain mostly or even exclusively written language, the number of spoken corpora based on both audio and video recordings has fortunately increased considerably over the last decade or so. This has complicated the process of annotation, given the many complexities that spoken, but not written, language from natural communicative settings implies. Most trivially, transcribers have to make choices regarding the orthographic representation of a spoken conversation with all its potential pitfalls: how to represent speech errors; pronunciations that differ from a standard dialect; how to represent a language for which there is no established writing system; whether or not to use capitalization and punctuation conventions, etc. But even if those problems are resolved, there are many other features of spoken language data that are worth annotating to facilitate corpus-linguistic research. These include, but are not limited to, phonological and prosodic characteristics, gestural and interactional and other characteristics as well as capturing the temporal quality of time series data and annotation.

2.2.1 Phonetic and phonological annotation

An orthographic transcription is the minimum requirement for a speech corpus, but a better representation of pronunciation may be desired for particular research questions. Speech may be annotated for phonemic transcription – that is, for the set of sounds that are phonemes in a language – or phonetic transcription, taking into account details of pronunciation. The former is usually considered to be broad in its detail, and a closed set of characters are usually used, though the set may be expanded to account for xenophones, sounds from other languages that may exist in borrowed words. In the past, annotators used a set of encoding 'hacks' to approximate the International Phonetic Alphabet, known as the Speech Assessment Methods Phonetic Alphabet (SAMPA; see Oostdijk & Boves 2008 for a history). With the growth of Unicode, however, the need for the SAMPA character set is obviated, although major corpora/resources like CELEX still use it.[2]

2 A reviewer points out that *entry* of IPA characters is still difficult on some computers, although software like IPA Palette (http://www.blugs.com/IPA/) make this task easier

Phonemic annotation is possible to generate automatically from orthographic transcription via a pronunciation lexicon and/or rule-based algorithms. Fine phonetic transcription, on the other hand, makes use of an extended set of characters including diacritics, and usually requires hand-coding by humans. Variations in pronunciation or certain kinds of allophony may be difficult to predict. Hand-coding is understandably expensive, and it is generally accepted that one minute of spoken language can require between 40 minutes and an hour to transcribe properly.

2.2.2 Prosodic annotation

Annotation of prosody occurs on a spectrum from broad, discourse-level prosodic generalization to detailed attention to small pitch changes across an utterance. Note that prosodically-annotated corpora are still not mainstream in corpus linguistics, and research on this (and other) paralinguistic aspects of speech is still in its early phases. As Oostdijk & Boves (2008:654) note,

> [b]ecause prosody constitutes a very important aspect of speech, one might expect that spoken language corpora come with some kind of prosodic annotation. Unfortunately, linguists do not agree on what a minimal theory-neutral prosodic annotation might or should contain.

An obvious early exception is the London-Lund Corpus of Spoken English, which was in turn derived from the Survey of English Usage and the Survey of Spoken English. This corpus marks basic prosodic features like tone units, prominent nuclei of units, length of pause and degrees of stress. This corpus is at the discursive end of the prosodic annotation spectrum. Other such systems include Discourse Transcription (DT; Du Bois et al., 1992) and the system used for Conversation Analysis (CA; see, e.g., Sacks, Schegloff, & Jefferson 1974; Schegloff 2007).

DT was developed as a system for divorcing transcription from traditional grammatical structure and instead allowing prosodic units, here called intonation units, to be the basic unit of transcription and analysis of spoken language. The system includes some information about intonational contour at the end of units, primary and secondary accent (akin to phrase-level stress), as well as other vocal and nonvocal characteristics of a given sample of naturalist speech like coughing, pauses, and vox. The Santa Barbara Corpus of Spoken American English is the largest published corpus using the Du Bois et al. system. The CA system also attends to discourse-level prosodic phenomena, but while DT is primarily prosodic in intention, CA is generally considered to be concerned with research on interaction between discourse participants, and is thus discussed more below.

At the other end of the spectrum we find systems like ToBI (TOne and Break Indices), which aims to capture syllable-by-syllable variations in pitch. The system is designed to facilitate research on the Autosegmental-Metrical model of intonation phonological theory (e.g. Bruce 1977, Pierrehumbert 1980). ToBI includes four tiers of transcription: words, tones, break indices, and notes. The Tones tier use a system of H (high), L (low), and diacritic notations for capturing tonal phrase accents, boundary tones, downstep, etc. The Break Indices tier uses a numerical scale of 0-4 to indicate the relative weakness or strength of a tonal break between syllables, which in turn indicates the boundaries of intonational units. ToBI has been applied to many languages; see Jun (2005) for an overview.

The advent of extremely large multimodal corpora such as the corpus created through the Human Speechome Project (90,000 hours of video and 140,000 hours of audio recordings) takes the problems of dealing with audio and video to another level altogether, requiring the

---

than it has been.

development of new kinds of tools to manage the extraordinary amount of data involved (Roy 2009).

2.2.3 Sign language and gesture annotation

Nonverbal language and nonverbal aspects of spoken language can also be annotated. The creation of annotated video-based sign language corpora has been increasing drastically in the last decades, especially with the development of software to time-align annotation and video media. The DGS-Korpus Sign Language Corpora Survey (2012) lists 36 corpora for 17 sign languages in various states of completion. These include Sign Languages from a range of European nations (Germany, France, Spain, the Netherlands, Austria, Great Britain, Sweden, Denmark, Ireland, and Iceland), as well as American, Australian, New Zealand, Korean, Mali, and Benkala Sign Languages. Of the 31 of these that are at least partially annotated, most are annotated primarily for gloss, with a few also using the Hamburg Phonetic Notation System ("HamNoSys", Hanke 2004), a phonetic system in use since the 1990s, for a basic transcription. 14 of these corpora are lemmatized. Other annotations include tagging for mouthings, facial expression, deviations from citation form, direction and orientation, mime, role shift, non-manuals, head shakes, eye gaze, eye aperture, eye brow, gesture, cheeks, comments, translations, lexematic units, semantic categories, semantic role, spatial modification, clause boundaries, pointing, and part of speech. 24 of these corpora have annotations time-aligned to video, most using the software tools ELAN (Max Planck Institute 2014; Slotje & Wittenburg 2006) or iLex (University of Hamburg 2014).

A particularly rich example of a sign language corpus is the Auslan corpus, which contains 300 hours of video recordings of naturalistic and elicited Australian sign language from 256 participants edited down to approximately 150 hours of usable language production. Recordings are linked to annotation and metadata files; the annotation of (part of) the corpus includes basic sign tokens as well as literal translations, eyegaze direction, palm orientation, handshape, verb type, spatial modification and aspect marking of verbs, clause boundaries, argument type and semantic roles of participants. (Johnson 2013).

Another nonverbal, paralinguistic feature for annotation is gesture. While minimal gesture tagging may be included in finer levels of transcription in, say, the Du Bois et. al system, more recently researchers have attempted to focus on the explicit annotation of gesture in video corpora. Kipp et al. (2007) proposes a grid for annotating the temporal quality of gesture. The top tier of the grid is for gesture phases, which come in a predictable order and are annotated as such (preparation, hold, stroke, hold, retraction). Aligned to this tier is another tier for gesture phrases, which describe gesture shape and motion in terms of a simplified set of lexemes (e.g., the gesture of the "Calm" lexeme is defined as "gently pressing downward, palms pointing downward", p. 334). A final aligned tier groups phases and phrases into gesture units, or periods of gesture between periods of rest. This last tier contains a description of the nature of the at-rest period at the end of the unit (e.g. "at-side," "folded," etc.). Other parameters for describing gesture in the Kipp et al. system include hand height, distance of hand from body, radial orientation to the central axis of the speaker, and arm swivel.

There is no single agreed-upon method for annotating gesture, however. Another example is that of the Bielefeld Speech and Gesture Alignment Corpus (SaGA, Lücking et al. 2010), which tags the co-occurrence of speech and gesture to provide a basis for studying the nonlinguistic aspects of communication. This project focuses on the annotation of the stroke phase, which is annotated in SaGA along eight parameters, adapted from earlier work by Müller (1998), Kendon (2004), and Streeck (2008): indexing/pointing, placing an imaginary object, (an object is placed or set down within gesture space), shaping or sculpting an object with the hands,

drawing the contour of an object, posturing or using the hands to stand for a static representation of an object, indicating sizes or distances, iconically counting items, and hedging via "wiggling or shrugging" (Lücking et al 2010:93).

2.2.4 Interactional annotation

By far the most common kind of annotation of interactional features of discourse is the Conversation Analysis (CA) system. The system, first compiled by Jefferson (1978, 1983a, 1983b, 1985, 1996), uses a series of symbols to indicate various features of dialog. These include temporality or sequentiality of utterances (square brackets for overlapping speech between multiple participants, line numbers to indicate order of utterance); the presence and length of pauses (measured in tenths of a second); some intonational qualities including pitch rise or fall, nonphonemically lengthened segments, stress/emphasis; audible aspiration; unusually slow or fast pacing; disfluencies (*uh*, *uhm*); etc. (Schegloff 2007). Unlike Du Bois et al.'s Discourse Transcription, in which prosodic units form the basis of the system with the goal of studying grammar in discourse, the basic unit in CA is the turn-at-talk, with the goal of studying interaction and sequence between speakers engaged in discourse.

2.3 Other

Given the many different applications for which corpora have been studied, there is of course a large number of other annotation formats that are used. For lack of space, we cannot discuss many more, but instead focus somewhat broadly on three additional formats below and refer the reader to Garside, Leech, & McEnery (1997), Beal, Corrigan, & Moisl (2007a, b), and Lüdeling & Kytö (2008) for more discussion.

2.3.1 Multilingual corpora: parallel corpora and interlinearized glossed text

Annotation can include a translational equivalent into another language. Parallel corpora contain translations of texts in a source language into one of more other languages, with the translated elements linked or aligned across languages in units consisting of words, phrases, or sentences. These corpora may also contain other kinds of annotation, like part-of-speech tagging, or links to a time code in a corresponding media file. In corpus linguistics, parallel corpora are usually smaller and more limited in genre than a single-language written corpus (Aijmer 2008), but are usually in larger, national languages, especially European languages, for which the European Union plays a large role in motivating the creation of parallel corpora (such as the European Parliament Proceedings Parallel Corpus; cf. Koehn 2005).

Documentary linguistic corpora are not usually thought of as "parallel corpora," but that is essentially what they are. Corpora of smaller, understudied languages often contain materials that have been annotated for translation on several levels. These are usually referred to as interlinearized glossed texts (IGT) and usually contain translations from the language of study to a language of greater communication (e.g. English) at the level of the morpheme, the word, and/or the phrase. IGT may contain other kinds of annotation as well, such as part of speech tagging, grammatical or constituency analysis, and prosodic information. The use of multilingual corpora extends from machine translation and language engineering, to translation studies, to lexicography, to the study of grammatical or typological phenomena.

2.3.2 Learner corpora

The last 10-15 years have seen a rapid increase in learner corpus research, i.e. corpus-based research on non-native language use by second/third/foreign language learners. This development has been facilitated by a variety of corpus compilation project, most notably the

International Corpus of Learner English (ICLE), under the leadership of the Centre for English Corpus Linguistics at the Université Catholique de Louvain. Learner corpora pose challenges to endeavors to annotate corpora, in particular to attempts at automatic annotation, given the fact that non-native language use is more likely than (edited) native language use to contain non-standard spellings, lexical items, and grammatical constructions that training data for, say, native-language lemmatizers, part-of-speech taggers, and parsers are unlikely to contain. Thus, such annotation efforts will likely require great care in choosing the right tagset and tagging algorithm (cf. van Rooy & Schäfer 2002), and more manual checking than is customary for native language use. One learner corpus project for which English is not the target language is the Corpus of Taiwanese Learners' Corpus of Spanish, which contains data from Taiwanese speakers (L2: English, L3: Spanish) of different levels from 15 universities. The corpus is richly annotated for parts-of-speech, lemmas, and errors made by the learners, and made available in XML format (Lu 2010).

The kind of annotation that is most naturally connected to learner corpora is error annotation, i.e. the identification of non-standard/non-native linguistic expressions in the learner data. Errors are usually annotated with regard to what would seem to be the target expression a native speaker would have produced in the identical context. Here, too, a fully automatic annotation process is not likely to succeed, which is why error annotation is usually done in a computer-assisted or even entirely manual fashion. The best-known error tagger is the Louvain error tagger, which assigns altogether 43 error tags, 31 in the categories of lexis, grammar, and lexico-grammar and 12 in the categories of form, punctuation, register, style, and word redundancies/omissions/ordering, but a variety other semi-automatic taggers have been used more narrowly too. Given the recency of these developments, the diversity of the tag sets employed in different projects, and the lack of availability of several error taggers for comparison, it is difficult to evaluate the degree of progress in the field of computer-aided error analysis, but it is clear at this point that the most important areas for further developments are standardization of tagsets both within and across target languages and automatization; cf. Díaz-Negrillo (2007: Section 2.5).

2.3.3 Discourse-pragmatic annotation

A still relatively rare but growing form of annotation encodes discourse-pragmatic information in texts. It is probably fair to say, however, that this annotation has mostly been applied in computational linguistics / natural language processing setting rather than in corpus linguistics proper, which is why we do not discuss this in depth. Examples for such corpora include the Lancaster Anaphoric Treebank, the Rhetorical Structure Discourse Treebank (Carlson, Marcu, & Okurowski 2003), which contains, "among other data, […] articles from the Penn Treebank, which were annotated with discourse structure in the framework of Rhetorical Structure Theory" (Zinsmeister et al. 2008:762), the EUSKAL RST Treebank-A (<https://ixa.si.ehu.es/Ixa/resources/Euskal_RSTTreebank>), a very small corpus (approximately 3K words) of abstracts of medical articles annotated on the basis of Rhetorical Structure Theory (Iruskieta, Diaz de Ilarraza, & Lersundi to appear), and the Penn Discourse Treebank (Prasad et al. 2008). Mitkov (2008) briefly discusses examples of bi-/multilingual parallel corpora which have been annotated for anaphoric or coreferential relationships; cf. Garside, Fligelstone, & Botley (1997) and Mitkov (2008) for much more information as well as discussion of how to assess inter-annotator agreement.

In addition to the above, corpora may also feature what is called pragmatic annotation. However, given the difficulty of even clearly defining what pragmatics per se is, it comes as no surprise that very many kinds of pragmatic annotation are conceivable. Archer, Culpeper, &

Davies (2008) (cf. also Leech, McEnery, & Wynne 1997) distinguish the annotation of formal components (based on words' and constructions' inherently pragmatic meaning), illocutionary force/speech, inferences (from Gricean maxims), interactional features above and beyond those discussed in Section 2.2.4, and various types of contextual information (linguistic and physical contexts, social, cultural, and cognitive contexts, etc.).

Finally, as an example of a corpus that combines very many kinds of annotation, consider The Narrative Corpus, which contains more than 500 narratives, socially balanced in terms of participant sex, age, and social class that were extracted from the demographically-sampled subcorpus of the British National Corpus. It contains sociological and sociolinguistic information on the speakers represented in the corpus, titles, subgenres, and textual components of the narratives, pragmatic and stylistic characteristics of the utterances (e.g., narrator and recipient roles or presentation modes), which are provided as inline XML annotation integrated with the existing BNC XML annotation (cf. Rühlemann & Brook O'Donnell to appear).

3 How are corpora annotated and exploited

That machine readability and interoperability requires some degree of standardization of annotation is somewhat of a truism in contemporary corpus linguistics; nonetheless, here we discuss two important aspects of annotation standardization: the use of Unicode, and the use of XML.

Unicode is a font-independent system for character encoding to ensure readability across languages and scripts. The Unicode Consortium publishes *The Unicode Standard* and a series of code charts; Unicode-enabled software can thus properly recognize and render (given the presence of an appropriate font) any Unicode character based on its underlying codepoint. For example, if a corpus creator renders the IPA character known as "voiceless retroflex plosive" (found in Hindi among other languages) with the Unicode code point 0288, any Unicode-enabled software will properly render this as ʈ. The importance of Unicode to corpus linguistic is obvious, as researchers can theoretically use any Unicode corpus in combination with any other.

Fortunately, another standard used in much of corpus linguistics already promotes the use of Unicode: XML. XML stands for eXtensible Markup Language, and is a language used for storing and transporting data based on its inherent structure (see Carletta et al. 2004). Elements in a given body of data are marked with a set of customizable tags which can be further defined using attributes. Elements are embeddable inside other elements as the data structure warrants (for example, "word" elements can be embedded inside "sentence" elements). XML has the advantage of being human-readable, but it must adhere to proper syntax, and tags and attributes must be defined in a separate document called a Document Type Declaration or a Schema.

Data properly stored in XML format can be easily converted into other formats (e.g., data bases) and for other uses via the use of a script designed to collect tagged elements as necessary. Thus a corpus properly tagged with valid XML can be searched and displayed. While XML is extensible, most corpus linguists will not need to write their own schema; there are already several standard versions of XML in use for corpus linguistics, including the Text Encoding Initiative (TEI, SOURCE), the EAF format used by ELAN annotation software, and Corpus Encoding Standard (XCES). Several XML metadata standards can also be used for corpora, including Dublin Core, Open Language Archives Community.

Several different kinds of annotation formats must be distinguished. First, the most frequent format is what is called *inline or embedded annotation*. In this format, which is heavily used for lemmatization and part-of-speech tagging, the annotation of a corpus file exists in the

same file and in the same line as the primary corpus data being annotated (and often comes in the form of SGML/XML annotation); we show multiple examples of this in Section 3.1. A sub-type of this annotation format is often used for parsed corpora, in which sentences are not shown with all words in one line as in the prototypical inline format, but are broken up across several lines to better show levels of syntactic embedding in parse trees to human users; examples are shown in Section 3.2.

Second, in *multi-tiered or interlinear annotation*, the primary corpus data and the annotation are in the same file but in different lines; more specifically, the primary corpus data are provided on separate lines from their annotations; one version of this format, CHAT, is particularly frequent in language acquisition corpora. Interlinearized glossed text, common to documentary corpora, is another popular format that is exemplified in Section 3.4. Note that multi-tiered annotation can also be easily converted to XML format for interoperability.

Finally, there are formats in which the primary corpus data and its annotation are stored in separate files or data structures. Such formats arise either from the storage of a corpus in a *relational database,* in which scholars provide limited but rapid search access to corpora via a website (e.g., <http://corpus.byu.edu/>) or, more usefully for more customizable and comprehensive access, when corpora come with so-called *standoff/standalone annotation*, in which the primary corpus data and their annotation are stored in separate (typically SGML/XML) documents linked to each other with hypertext (cf. Thompson & McKelvie 1997). While the corpus-as-database approach has become more frequent over the past 10 years, standoff annotation is unfortunately still rare in spite of its many advantages:

- "the base document may be read-only and/or very large, so copying it to introduce markup may be unacceptable;
- the markup may include multiple overlapping hierarchies;
- it may be desirable to associate alternative annotations (e.g., part-of-speech annotation using several different schemes, or representing different phases of analysis) with the base document;
- it avoids the creation of potentially unwieldy documents;
- distribution of the base document may be controlled, but the markup is freely available." (Ide 1998)

However, not all levels of annotation lend themselves equally easily to standalone annotation (see McEnery, Xiao, & Tono 2006:44), and at present very few tools for exploring corpora with standalone annotation are available: inline/embedded annotation can be handled somewhat satisfactorily with some of the most frequently-used ready-made software tools (e.g., AntConc, Anthony 2014) and very well with programming languages like R, Python, or Perl whereas standalone annotation is more challenging to explore (Zinsmeister et al. 2008:769).

3.1 Part-of-speech tagging (inline/embedded)

As mentioned above, the most frequent annotation is part-of-speech tagging, which is so prevalent because of the relative ease of annotation (especially in the languages for which many (large) corpora are available) and because many other forms of annotation require it to be present. In this subsection, we exemplify several of the most frequent POS-tagging formats. Figure 1 represents the first sentence of the Brown corpus of written American English without annotation (for comparison) while Figure 2 and Figure 3 represent the same sentence in different POS-tagging formats.

```
A01 0010    The Fulton County Grand Jury said Friday an investigation
```

```
A01 0020 of Atlanta's recent primary election produced "no evidence" that
A01 0030 any irregularities took place.   The jury further said in term-end
```

Figure 1: Brown corpus, simplest legacy version, sentence 1

```
|SA01:1 the_AT Fulton_NP County_NN Grand_JJ Jury_NN said_VBD Friday_NR an_AT investigation_NN
of_IN Atlanta's_NP$ recent_JJ primary_NN election_NN produced_VBD no_AT evidence_NN that_CS
any_DTI irregularities_NNS took_VBD place_NN ._.
```

Figure 2: Brown corpus, part-of-speech tagged, sentence 1

```
<p><s n="1">
        <w type="at">The</w>
        <w type="np-tl">Fulton</w>
        <w type="nn-tl">County</w>
        <w type="jj-tl">Grand</w>
        <w type="nn-tl">Jury</w>
        <w type="vbd">said</w>
        <w type="nr">Friday</w>
        <w type="at">an</w>
        <w type="nn">investigation</w>
        <w type="in">of</w>
        <w type="np$">Atlanta's</w>
        <w type="jj">recent</w>
        <w type="nn">primary</w>
        <w type="nn">election</w>
        <w type="vbd">produced</w>
        <c type="pct">``</c>
        <w type="at">no</w>
        <w type="nn">evidence</w>
        <c type="pct">''</c>
        <w type="cs">that</w>
        <w type="dti">any</w>
        <w type="nns">irregularities</w>
        <w type="vbd">took</w>
        <w type="nn">place</w>
        <c type="pct">.</c>
</s> </p>
```

Figure 3: Brown corpus, XML part-of-speech tagged, sentence 1

For English corpora, the most widespread part-of-speech tagsets are CLAWS (Constituent Likelihood Automatic Word-tagging System) C5 and C7. The former has 63 simple tags, the latter uses 137 word tags and additional punctuation mark tags. Figure 4 shows the POS-tagging of the BNC World Edition in SGML format whereas Figure 5 shows the same sentence in the XML annotation that is now standard; note how the latter provides a more explicit annotation to highlight the fact that *sort of* is treated as a multi-word unit (hence the <mw> tag) consisting of *sort* (NN1, a noun in the singular) and *of* (PRF).

```
<s n="1">
        <w VVB>Introduce            <w NP0>Brenda          <w PNQ>who<w VBZ>'s
        <w VVG>going                <w TO0>to              <w VVI>speak
        <w PRP>to                   <w PNP>us              <w AVP-PRP>on
        <w VVB>Make                 <w VDI>do              <w CJC>and
        <w VVB>Mend                 <w CJC>and             <w PNP>she
        <w VHZ>'s                   <w VVN>asked           <w PNP>me
        <w TO0>to                   <w VVI>say             <w CJT>that
        <w PNP>she                  <w VM0>'d              <w VBI>be
        <w AV0>very                 <w AJ0>pleased         <w CJS>if
        <w NN0>people               <w VVB-NN1>break       <w AVP>in
        <w CJC>or                   <w UNC>erm             <w AV0>sort of
        <w VVB-NN1>form             <w DT0>some            <w NN1>sort
        <w PRF>of                   <w NN1>dialogue        <w PRP>with
        <w PNP>her                  <w CJS>as              <w PNP>she
        <w VVZ>goes                 <w AVP>along           <c PUN>.
```

Figure 4: BNCwe SGML: D8Y, sentence 1

```
<s n="1">
        <w c5="VVB" hw="introduce" pos="VERB">Introduce </w>
        <w c5="NP0" hw="brenda" pos="SUBST">Brenda </w>
        <w c5="PNQ" hw="who" pos="PRON">who</w>
        <w c5="VBZ" hw="be" pos="VERB">'s </w>
        <w c5="VVG" hw="go" pos="VERB">going </w>
        <w c5="TO0" hw="to" pos="PREP">to </w>
        <w c5="VVI" hw="speak" pos="VERB">speak </w>
        <w c5="PRP" hw="to" pos="PREP">to </w>
        <w c5="PNP" hw="we" pos="PRON">us </w>
        <w c5="AVP-PRP" hw="on" pos="ADV">on </w>
        <w c5="VVB" hw="make" pos="VERB">Make </w>
        <w c5="VDI" hw="do" pos="VERB">do </w>
        <w c5="CJC" hw="and" pos="CONJ">and </w>
        <w c5="VVB" hw="mend" pos="VERB">Mend </w>
        <w c5="CJC" hw="and" pos="CONJ">and </w>
        <w c5="PNP" hw="she" pos="PRON">she</w>
        <w c5="VHZ" hw="have" pos="VERB">'s </w>
        <w c5="VVN" hw="ask" pos="VERB">asked </w>
        <w c5="PNP" hw="i" pos="PRON">me </w>
        <w c5="TO0" hw="to" pos="PREP">to </w>
        <w c5="VVI" hw="say" pos="VERB">say </w>
        <w c5="CJT" hw="that" pos="CONJ">that </w>
        <w c5="PNP" hw="she" pos="PRON">she</w>
        <w c5="VM0" hw="would" pos="VERB">'d </w>
        <w c5="VBI" hw="be" pos="VERB">be </w>
        <w c5="AV0" hw="very" pos="ADV">very </w>
        <w c5="AJ0" hw="pleased" pos="ADJ">pleased </w>
        <w c5="CJS" hw="if" pos="CONJ">if </w>
        <w c5="NN0" hw="people" pos="SUBST">people </w>
        <w c5="VVB-NN1" hw="break" pos="VERB">break </w>
        <w c5="AVP" hw="in" pos="ADV">in </w>
        <w c5="CJC" hw="or" pos="CONJ">or </w>
        <w c5="UNC" hw="erm" pos="UNC">erm </w>
        <mw c5="AV0">
                <w c5="NN1" hw="sort" pos="SUBST">sort </w>
                <w c5="PRF" hw="of" pos="PREP">of </w>
        </mw>
        <w c5="VVB-NN1" hw="form" pos="VERB">form </w>
        <w c5="DT0" hw="some" pos="ADJ">some </w>
        <w c5="NN1" hw="sort" pos="SUBST">sort </w>
        <w c5="PRF" hw="of" pos="PREP">of </w>
        <w c5="NN1" hw="dialogue" pos="SUBST">dialogue </w>
        <w c5="PRP" hw="with" pos="PREP">with </w>
        <w c5="PNP" hw="she" pos="PRON">her </w>
        <w c5="CJS" hw="as" pos="CONJ">as </w>
        <w c5="PNP" hw="she" pos="PRON">she </w>
        <w c5="VVZ" hw="go" pos="VERB">goes </w>
        <w c5="AVP" hw="along" pos="ADV">along</w>
        <c c5="PUN">.</c>
</s>
```

Figure 5: BNCwe XML: D8Y, sentence 1

As is seen from the above, this kind of annotation of the BNC World Edition also includes lemmatization (hw="…") and major parts of speech (pos="…"), which means that quite comprehensive searches can be performed.

Most of the time, part-of-speech annotation is provided inline/embedded as in all of the above examples. The American National Corpus Open is available in the XML form represented in Figure 6, which also contains annotation for syntactically-informed noun chunks, as well as in a format called standoff/standalone annotation, in which primary data and (different layers of) annotation are stored in separate files that are linked together by pointers.

```
<turn id="t32" who="EA">
        <u id="t32u1"><u id="t32u1">
                <NounChunk><tok base="i" msd="PRP">I</tok></NounChunk>
                <tok base="pretty" msd="RB">pretty</tok>
                <NounChunk>
                        <tok base="much" msd="JJ">much</tok>
                        <tok base="remember" msd="VB">
```

```
                        <VG tense="Inf" type="NFVG"
                        voice="active">remember</VG></tok>
                    <tok base="the" msd="DT">the</tok>
                    <tok base="whole" msd="JJ">whole</tok>
                    <tok base="thing" msd="NN">thing</tok>
                </NounChunk>
                <tok base="." msd=".">.</tok>
        </u></u>
</turn>
```

Figure 6: ANC Open: AdamsElissa, line 150-152

3.2 Parsed corpora (inline/embedded)

In this section, we briefly exemplify syntactic parsing in corpora. Figure 7 exemplifies parsing as used in the British Component of the International Corpus of English, which contains POS-tags and also a parse tree (with all words in curly brackets and whitespace indentation reflecting the depths of branching.

```
[<#3:1:A> <sent>]
PU,CL(main,inter,intr,past)
 DISMK,FRM {Sorry}
 INTOP,AUX(modal,past) {could}
 SU,NP()
  NPHD,PRON(pers) {you}
 VB,VP(intr,infin,modal)
  MVB,V(intr,infin) {start}
 A,AVP(ge)
  AVHD,ADV(ge) {again}
[<$B>]
```

Figure 7: ICE-GB S1A-001, parse unit 3

Figure 8 is an example of the widely used Penn Treebank annotation

```
( (S (NP-SBJ-1 Jones)
     (VP followed
         (NP him)
         (PP-DIR into
                 (NP the front room))
         ,
         (S-ADV (NP-SBJ *-1)
                 (VP closing
                     (NP the door)
                     (PP behind
                         (NP him)))))
     .))
```

Figure 8: Example of Penn Treebank annotation (from Taylor, Marcus, & Santorini 2003:10)

Some parsed corpora are provided in yet different formats. An example is the NEGRA Corpus, a parsed corpus of German newspaper texts (355K words, 20.6K sentences), which are available both in the Penn Treebank format and in an export format exemplified in Figure 9.

```
%% word                  tag     morph           edge    parent  secedge comment
#BOS 2 2 899973978 1
Sie                     PPER    3.Pl.*.Nom      SB      504
gehen                   VVFIN   3.Pl.Pres.Ind   HD      504
gewagte                 ADJA    Pos.*.Akk.Pl.St NK      500
Verbindungen            NN      Fem.Akk.Pl.*    NK      500
und                     KON     --              CD      502
Risiken                 NN      Neut.Akk.Pl.*   CJ      502
ein                     PTKVZ   --              SVP     504
,                       $,      --              --      0
```

```
versuchen               VVFIN   3.Pl.Pres.Ind   HD      505
ihre                    PPOSAT  *.Akk.Pl        NK      501
Möglichkeiten           NN      Fem.Akk.Pl.*    NK      501
auszureizen             VVIZU   --              HD      503
.                       $.      --              --      0
#500                    NP      --              CJ      502
#501                    NP      --              OA      503
#502                    CNP     --              OA      504
#503                    VP      --              OC      505
#504                    S       --              CJ      506
#505                    S       --              CJ      506
#506                    CS      --              --      0
#EOS 2
#BOS 3 2 916759524 1
```

Figure 9: Export annotation format of the NEGRA corpus

Finally, as an example for a dependency-based treebank, consider Figure 10 for the Reference Corpus for the Processing of Basque (EPEC; cf. Aldebazal et al. 2009), a 300K word corpus of written Basque annotated morphologically (for part-of-speech, number, definiteness, and case), lexically (for named entities, multi-word units), and syntactically in a Dependency-Grammar format.

(9) *Euri-ak* *ez* *zaitu* *bustitzen* *Valentine*
Rain-SG-ERG not AUX-PRS-2SG-3SG wet-IPFV Valentine-VOC
'The rain is not wetting you, Valentine.'

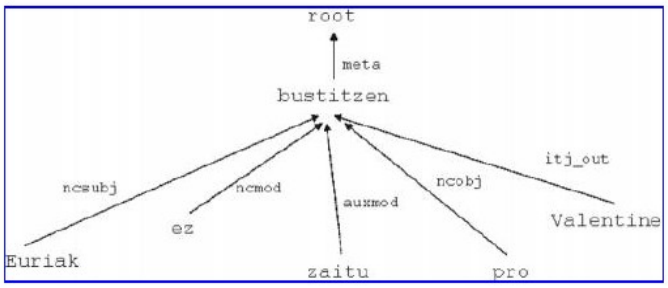

Figure 4. *Dependency tree for* Euriak ez zaitu bustitzen, Valentine

Figure 10: Example of EPEC annotation (Aldebazal et al. 2009:255)

3.3 Other annotation (inline/embedded)

In this section, we exemplify a few other, less widely used formats of inline/embedded annotation. Figure 11 is a brief example of the semantic-annotation format used in ProbBank (cf. Section 2.1.4 above).

```
[ARGM-LOC  In such an environment] , [ARG0  a market maker]
[ARG-MOD  can]  [rel  absorb]  [ARG1  huge losses] .
```

Figure 11: Example of PropBank annotation (from Zinsmeister et al. 2008:762)

Figure 12 shows error annotation in learner corpora: errors are marked with letter sequences in parentheses preceding an error (FS = form + spelling, GADJN = grammar + adjective + number, etc.) and intended targets in $ signs following an error.

```
There was a forest with dark green dense foliage and pastures where a herd of tiny
```

```
(FS) braun $brown$ cows was grazing quietly, (XVPR) watching at $watching$ the toy
train going past. I lay down (LS) in $on$ the moss, among the wild flowers, and looked
at the grey and green (LS) mounts $mountains$. At the top of the (LS) stiffest
$steepest$ escarpments, big ruined walls stood (WM) 0 $rising$ towards the sky. I
thought about the (GADJN) brutals $brutal$ barons that (GVT) lived $had lived$ in
those (FS) castels $castles$. I closed my eyes and saw the troops observing (FS)
eachother $each other$ with hostility from two (FS) opposit $opposite$ hills.
```

Figure 12: Sample of error-tagged text (Dagneaux, Denness, & Granger 1998:16, quoted from Díaz-Negrillo 2007:62f.)

Transcription of spoken language presents considerable challenges, at least if one wishes to highlight faithfully features particular to spoken language like overlapping speech. The annotated transcription in Figure 13, a sample of transcribed spoken language taken from ICE-CAN, illustrates some of this complexity. Overlapping strings are indicated by <[>…</[>, with the complete set of overlapping strings contained within <{>…</{>, stretching across both speaker A and speaker B. The tags <}>…</}> indicate a "normative replacement," where a repetition of *they* (in casual, face-to-face conversation) is indicated. This annotation allows for searching on the raw data (containing the original two instances of *they*) or on the normalized version (containing one instance of *they* within <=…></=>).

```
<$A> <ICE-CAN:S1A-001#34:1:A> I think some of the trippers actually do a bit of the
portaging by themselves <}> <-> they> </-> <=> they </=> </}> bring it to the other
end and they come back to help the kids with <{> <[> their packs </[>
<$B> <ICE-CAN:S1A-001#35:1:B> <[> I see </[> </{>
```

Figure 13: Overlap marking from ICE-CAN S1A-001

Finally, Figure 14 is an example of discourse-pragmatic annotation showing the UCREL scheme annotation for cohesive relationships, where the antecedent NP *Kurt Thomas* is parenthesized and numbered and then referred back to with <. While this annotation format does not use standardized SGML/XML annotation, later developments for anaphoric-relations tagging, such as the MUC annotation scheme (Hirschmann & Chinchor 1997, are SGML-based and, thus, allow for easier exchange of data and results.

```
Anything (108 Kurt Thomas 108) does, <REF=108 he does to win. Finishing
second, <REF=108 he says is like finishing last.
```

Figure 14: Example of the UCREL annotation (from Mitkov 2008:584; cf. also Garside, Fligelstone, & Botley 1997 for details)

3.4 Multi-tiered and other annotation

Multi-tiered annotation is a method of displaying and structuring data that assumes a relationship between items shown on different tiers or lines. Interlinearized Glossed Text (IGT) is an example of multi-tiered annotation that has traditionally been a display format for segmented samples of speech and translating them into another language, as shown in Figure 15:

Aka faupuskam munaa uɾi.
a=ka fau-pus-ka-m muna=a uɾi
I=TOP eat-DES-VBZ-IND thing=TOP PROX
'That's what I want to eat.'

Figure 15: Example of IGT in Ōgami (Miyako Ryukuyan), (Pellard 2010:153.)

While the relationship between tiers may not be explicitly marked, a range of information can be gleaned from the layout of the IGT. Morpheme borders are indicated in the second line, as well as the category of morpheme: affixes are marked with hyphens, and clitics are marked with equal signs. Word boundaries are marked with whitespace. Glosses are given at the morpheme level in line 3 and are aligned to the left edge of the word. Although this example does not overtly align morphemes with their glosses, this information can be deduced by counting morpheme boundaries (and there is no reason why one could not also align morphemes to their glosses). Grammatical category information is also given in line 3, with lexical items glossed in plain type and grammatical morphemes glossed in small caps. A part of speech line could be added if desired. The entire sentences is aligned to its free translation into English, shown in line 4.

However, in the past IGT was simply a method for printed display, and not necessarily structured in a way that made machine reading possible. Advances in tools such as Toolbox give structure to IGT by using "backslash codes" known as Multi-Dictionary Format (MDF) tags, as in Figure 16. The MDF tags at the beginning of each line indicate the content contained there, in a hierarchical relationship with \id, the parent tag in this example. The item with the identification number 061:005 has corresponding audio (\aud), a line of transcription (\tx), a morphemic parse (\mr), a morphemic gloss (\mg), and a free gloss (\fg). MDF contains many more backslash codes for lexical tagging.

\id	061:005							
\aud	AHT-MP-20100305-Session.wav 02:19.320-02:21.780							
\tx	Ga łdu' ben yii taghił'aa.							
\mr	ga	łdu'	ben	yii	ta-	ghi-	ł-	'aa
\mg	DEM	FOC	lake	in	water	ASP	CLF	linear.extend
\fg	'As for that one (river), it flows into the lake.'							

Figure 16: Example of Toolbox format of IGT, showing MDF tags (Thieberger & Berez 2012:96)

Another example of an attempt to make structural relationships between tiers explicit is the very widely used CHAT format as shown in Figure 17 below.

```
*CHI:   more cookie . [+ IMP]
%mor:   qn|more n|cookie .
%gra:   1|2|QUANT 2|0|ROOT 3|2|PUNCT
%int:   distinctive, loud
%trn:   qn|more n|cookie .
%gra:   1|2|QUANT 2|0|ROOT 3|2|PUNCT
```

Figure 17: CHAT format annotation from CHILDES data (Brown: Eve01.cha, utterance 1)

Here tier labels perform the function of indicating the relationship between the child's utterance (labeled *CHI) and the various types of annotation: morphemic analysis (%mor), grammatical relations (%gra), intonation (%int), a hand-annotated version of the %mor tier for training/checking (%trn), and many others allowing to annotate nearly all of the types of information discussed in Section 2 (action, addressees, cohesion, facial gestures, paralinguistic information, etc.).

The above is a legacy format which is mainly explored with a software called CLAN

(<http://childes.psy.cmu.edu/clan/>). CLAN is freely available for Windows, Mac, and Unix/Linux and allows the researcher to generate frequency lists, compute type-token ratios or more sophisticated measures of vocabulary richness/lexical diversity, generate concordances using regular expressions to retrieve lexical items, particular parts of speech (and their combinations), etc. However, one specific advantage of CLAN's handling of the annotation is how the user can return from textual results to the relevant audio or video.

However, over the last few years, XML versions of a large amount of the data in CHILDES have been made available, which can now be explored with more general and more powerful tools. Here's the above sentence from EVE01.cha in its XML form:

```
<u who="CHI" uID="u0">
  <w>more<mor type="mor"><mwg><mw><pos><c>qn</c></pos><stem>more</stem></mw></mwg></mor>
  <mor type="trn"><mwg><mw><pos><c>qn</c></pos><stem>more</stem></mw></mwg></mor></w>
  <w>cookie<mor type="mor"><mwg><mw><pos><c>n</c></pos><stem>cookie</stem></mw></mwg></mor>
  <mor type="trn"><mwg><mw><pos><c>n</c></pos><stem>cookie</stem></mw></mwg></mor></w>
  <t type="p"/>
  <postcode>IMP</postcode>

  <a type="extension" flavor="xgra">1|2|QUANT 2|0|ROOT 3|2|PUNCT</a>
  <a type="intonation">distinctive, loud</a>
  <a type="extension" flavor="xGRA">1|2|QUANT 2|0|ROOT 3|2|PUNCT</a>
</u>
```

Figure 18: XML annotation from CHILDES data (Brown: Eve01.cha, utterance 1)

A final example that combines the rarer cases of phonetic and non-inline annotation is the Up corpus based on the "Up" series of documentary films by director Michael Apted, containing data on a set of individuals at seven-year intervals over a period of 42 years and exemplified in Figure 19 representing the annotation of "give me" spoken by a male speaker.

> The corpus is meant to facilitate phonetic, psycholinguistic and sociolinguistic research on age-related change in speech during young and middle-age adulthood. The corpus contains audio files, transcripts time-aligned at the level of utterance, word, and segment, F0 and vowel formant measurements of portions of the films featuring eleven participants at ages 21 through 49. (Gahl to appear: abstract)

ptoken_id	phone	ptoken_start	ptoken_end	word	speaker	age	sex
346674	g	624.44	624.48	GIVE	nick	35	male
346675	ɪ	624.48	624.52	GIVE	nick	35	male
346676	v	624.52	624.58	GIVE	nick	35	male
346677	m	624.58	624.66	ME	nick	35	male
346678	i	624.66	624.71	ME	nick	35	male

Figure 19: Annotation in the "Up" Corpus

While the above discussion showcases quite a few formats, the more complex the annotation, the less straightforward it can be to exemplify; for example, standoff annotation is more difficult to visualize given how links between points in separate (XML) documents would have to be represented. This problem will be exacerbated even more in, for example, multimodal corpora. Multimodal corpora present challenges for mapping layers of annotation to time series data like audio and video recordings. Bird & Liberman (2001) present a model for the logical structure of layers of annotation and time known as an *annotation graph***.** An annotation graph allows for the flexible establishment of a hierarchical series of annotation nodes with a fundamental node based on either character position for text corpora or time offsets for speech

corpora. The graph can accommodate many kinds of annotation and logical structures, including orthographic and phonetic transcription, syntactic analysis, morphological analysis, gesture, part of speech, lemmatization, etc. Furthermore, the annotation graph allows the establishment of time-based events that overlap or gap, the division of those events into time-based or abstract subdivisions (e.g. time-alignment of words, or non-time-aligned morphemic parses respectively), as well as symbolically-related annotations like translations.

Although Zinsmeister (2008:767) was skeptical that the annotation graph could be made functional ("[...] it is difficult to imagine a general tool that would allow the user to access the whole range of annotations without having an overly complex and cryptic user interface"), ELAN is one annotation tool based on the annotation graph. Provided the user understands the data structure and the relationships between different layers of annotation and can map them onto one of the software's built-in models of data types, ELAN creates customizable and logically sound multi-layered annotation that is time-aligned to corresponding media. In any case, data in an XML instantiation of the annotation graph model can be exported to yield formats as those exemplified above as well as searched/processed via regular corpus linguistic methods for XML data.

4 Concluding remarks

While it cannot be denied that there are still some voices in corpus linguistics arguing against linguistic annotation – most notably the late John Sinclair and other scholars from the Birmingham-school inspired corpus-driven linguistics camp (cf, e.g., Hunston 2002) – linguistic annotation is here to stay: While annotation might in theory turn out to be distracting or misleading on occasion, obviously no corpus linguist is obligated to rely on, use, or even view the corpus annotation in a particular study. Thus, the majority view in contemporary corpus linguistics is that "adding annotation to a corpus is giving 'added value'" to it (Leech 2005: Section 1) and that explicit annotation of the type discussed in this volume is superior to the 'implicit annotation that results from "applying intuitions when classifying concordances […] which unconsciously makes use of preconceived theory", and which is "to all intents and purposes unrecoverable and thus more unreliable than explicit annotation." Xiao (2008:995). That is, annotation "only means undertaking and making explicit a linguistic analysis" (McEnery, Xiao, & Tono 2006:32).

As has become clear from even this cursory overview, multiple kinds of annotation are being used and the number of annotated resources that add value to primary data is steadily increasing; at the same time, there is a lot of work on the improvement of existing, and development of new, annotation formats that are bound to allow for ever more comprehensive searches and research. In this final section, we summarize a few desiderata for such work that can, hopefully, inspire new developments and renewed attention to problems that corpus linguists regularly face in their work.

Obviously, the raison d'être of annotation in general is to allow corpus linguists to retrieve all and only all instances of a particular phenomenon. Given the complexity and multi-layeredness of linguistic data, this leads to two main desiderata. One is that, as annotation for more and more subjective characteristics becomes more frequent, it is imperative that annotation provides efficient ways for dealing with ambiguous or otherwise problematic data points. In the comparatively simple domain of part-of-speech tagging, for example, this means finding efficient ways to deal with uncertainty in the assignment of tags: some tagsets use portmanteau tags that indicate that the tagger had insufficient evidence to make a clear distinction between two tags.

For example, in the BNC the form *spoken* may be annotated as <w AJ0-VVN> for 'adjective in the base form' or 'verb in the past participle') or in the Penn Treebank the form *entertaining* may be annotated as [JJ|VBG] for 'adjective' or verb in the 'gerund'. Similarly, annotation faces potentially difficult questions when it comes to tagging clitics such as *don't*. Those are annotated as <w VDB>do<w XX0>n't in the BNC SGML (VDB = 'base form of the verb *do*, XX0 = *not*/*n't*), which is compatible with do_DO n't_XNOT in the Lancaster-Oslo-Bergen corpus and an annotation of *innit* as <w VBZ>in<w XX0>n<w PNP>it, which at first sight may seem surprising (because the tag VBZ – third person singular of the verb *be* – is applied to what seems to be the preposition *in*, PNP = personal pronoun).

Other important questions arise with multiple layers of annotation. On the one hand, this may arise when there are different layers of annotation (either different tagsets for the same conceptual level such as part-of-speech tagging or different levels of annotation as when syntactic parsing and semantic annotation for one and the same corpus are to be combined); unfortunately, no definite best practices or standards seem to have emerged yet, given the recency and speed of new developments in annotation and tool development. On the other hand, annotation questions even arise in the seemingly much simpler process of tokenization of, say, multi-word units; recall how Figure 5 showed how multi-word units are annotated in the current version of the BNC World Edition (here repeated as Figure 20), which complicates retrieval processes with some widespread concordancing tools, and maybe even programming languages.

```
<mw c5="AV0">
        <w c5="NN1" hw="sort" pos="SUBST">sort </w>
        <w c5="PRF" hw="of" pos="PREP">of </w>
</mw>
```

Figure 20: Multi-word units in the BNC World Edition

Issues like these become even much more challenging once corpus linguists turn more from the currently prototypical corpora on the currently most-studied languages – the usual Indo-European suspects – to currently less frequent audio/multimodal corpora and corpora of (much) lesser-studied languages, whose morphosyntactic characteristics may require forms of annotation that go beyond what the field is presently accustomed to. Forays into corpus based methods in these languages have resulted in answers to longstanding linguistic questions that had remained unanswered via other methods (e.g. Berez & Gries 2010), and the goals of corpus linguistics and language documentation are not so different (Cox 2011, McEnery & Ostler 2000, Ostler 2008). Both fields aim for collections of related language data that are interoperable, searchable, reusable, and mobilizable for a broad range of linguistic inquiry. While corpus theorization and creation may be more limited for small or endangered languages – for example, balance and representativeness are often limited by the number and skill of available speakers – standards for annotation can, with more discussion between practitioners on both sides, become more broadly useful across disciplines. Current advances in encoding and interoperablility like XML and Unicode are already making this possible.

Most of these challenges are being addressed in various ways and can probably be handled extremely well with the kind of standoff annotation that has been recommended for more than a decade. However, as alluded to above, corpus linguistics is at an evolutionary and generation-changing moment. Many, if not most, practitioners are dependent on a very small set of ready-made (often proprietary) concordancing tools and the transition to a more wide-spread command of programming languages and regular expressions is only happening now (quite unlike in computational linguistics / natural language processing). Thus, while the field is increasingly 'demanding' more and more sophisticated corpora and annotations, technical skills

still need to evolve more to a point where the most recent developments in annotation can be utilized to their fullest. The really most central desiderata are therefore

- the development of corpus exploration tools that strike a delicate balance between facilitating the exploration of corpora that have been comprehensively annotated;
- continued research and development of tools that allow for reliable conversions of the many different annotation formats used by many different tools (cf. MacWhinney 2011:187);
- the continuing evolution of the field towards more technical skills/expertise and less dependence on two or three concordancing tools that do not provide the versatility that today's annotation complexity requires;
- the sharing of annotation practices and standards among corpus annotators working on small and large languages alike.

Only when all these desiderata are met will corpus linguistics as a discipline be able to take its research to the next evolutionary level.

References

Aijmer, Karin. 2008. Parallel and comparable corpora. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics: An international handbook*, 275-292. Berlin, New York: Walter de Gruyter.

Aldebazal, Izaskun, Maria Jesus Aranzabe, Jose Mari Arriola, & Arantza Dias de Ilarraza. 2009. Syntactic annotation in the Reference Corpus for the Processing of Basque (EPEC): theoretical and practical issues. *Corpus Linguistics and Linguistic Theory* 5(2). 241-269.

Anthony, Laurence. 2014. AntConc: a freeware concordance program for Windows, Macintosh OS X, and Linux. URL <http://www.antlab.sci.waseda.ac.jp/antconc_index.html>.

Archer, Dawn, Andrew Wilson, & Paul Rayson. 2002. Introduction to the USAS category system. Unpublished ms, Lancaster University <http://ucrel.lancs.ac.uk/usas/usas %20guide.pdf>.

Archer, Dawn, Jonathan Culpeper, & Matthew Davies. 2008. Pragmatic annotation. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics: An international handbook*, 613-642.Berlin, New York: Walter de Gruyter.

Bard, Ellen G., Catherine Sotillo, Anne H. Anderson, Henry S. Thompson, & M. Martin Taylor. 1996. The DCIEM Map Task Corpus: Spontaneous dialogue under sleep deprivation and drug treatment. *Speech Communication* 20 (1/2). 71-84.

Beal, Joan C., Karen P. Corrigan, & Hermann L. Moisl (eds.). 2007a. *Creating and digitizing language corpora. Vol. 1: Synchronic databases*. Houndmills: Palgrave Macmillan.

Beal, Joan C., Karen P. Corrigan, & Hermann L. Moisl (eds.). 2007b. *Creating and digitizing language corpora. Vol. 2: Diachronic databases*. Houndmills: Palgrave Macmillan.

Berez, Andrea L. & Stefan Th. Gries. 2010. Correlates to middle marking in Dena'ina iterative verbs. *International Journal of American Linguistics* 76(1):145-165.

Bird, Steven & Mark Liberman. 2001. A formal framework for linguistic annotation. *Speech Communication* 33(1-2). 23-60.

Brants, Sabine, Stefanie Dipper, Peter Eisenberg, Silvia Hansen, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, & Hans Uszkoreit. 2004. TIGER: Linguistic Interpretation of a German Corpus. *Journal of Language and Computation* 2004(2). 597-

620.
Carletta, Jean, David McKelvie, Amy Isard, Andreas Mengel, Marion Klein, & Morten Baun Møller. 2004. A generic approach to software support for linguistic annotation using XML. In Geoffrey Sampson & Diana McCarthy (eds.), *Corpus linguistics: readings in a widening discipline*, 449-459. London & New York: Continuum.
Cox, Christopher. 2011. Corpus linguistics and language documentation: challenges for collaboration. In John Newman, R. Harald Baayen, & Sally Rice (eds.), *Corpus-based studies in language use, language learning, and language documentation*, 239-264. Amsterdam: Rodopi.
Czaykowska-Higgins, Ewa. 2009. Research models, community engagement, and linguistic fieldwork: Reflections on working with Canadian Indigenous communities. *Language Documentation & Conservation* 3(1). 15-50.
Dagneaux, Estelle Sharon Denness, & Sylviane Granger. 1998. Computer-aided error analysis. *System* 26. 163-174.
DGS-Korpus Sign Language Corpora Survey. http://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/sl-corpora.html. Accessed September 20, 2013.
Díaz-Negrillo, Ana. 2007. A fine-grained error tagger for English learner corpora. Unpublished Ph.D. thesis, University of Jaén.
Du Bois, John W., Susanna Cumming, Stephan Schuetze-Coburn, & Danae Paolino (eds.). 1992. *Discourse transcription (Santa Barabara Papers in Linguistics, Volume 4)*. Santa Barbara: University of California, Santa Barbara Department of Linguistics.
Fellbaum, Christiane (ed.). 1998. *WordNet: An electronic lexical database*. Cambridge, MA: The MIT Press.
Fellbaum, Christiane, Joachim Garabowski, Shari Landes, & Andrea Baumann. 1998. Matching words to senses in WordNet: Naïve vs. expert differentiation. In Christiane Fellbaum (ed.). 1998. *WordNet: An electronic lexical database*, 217-239. Cambridge, MA: The MIT Press.
Fillmore, Charles J. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*. Volume 280. 20-32.
Fitschen, Arne & Piklu Gupta. 2008. Lemmatising and morphological tagging. In Anke Lüdeling & Merja Kytö (eds.). *Corpus linguistics: an international handbook. Vol. 1*, 552-564. Berlin & New York: Walter de Gruyter.
Gahl, Susanne. to appear. The "Up" Corpus: A corpus of speech samples across adulthood. *Corpus Linguistics and Linguistic Theory.*
Garofolo, John S., Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, & Victor Zue. 1993. *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Philadelphia: Linguistic Data Consortium.
Garside, Roger, Steve Fligelstone, & Simon Botley. 1997. Discourse annotation: anaphoric relations in corpora. In Roger Garside, Geoffrey Leech, & Tony McEnery (eds.). 1997. *Corpus annotation: linguistic information from computer text corpora*, 66-84. London: Longman
Garside, Roger, Geoffrey Leech, & Tony McEnery (eds.). 1997. *Corpus annotation: linguistic information from computer text corpora*. London: Longman.
Godfrey, John J. & Edward Holliman. 1997. *Switchboard-1 Release 2*. Philadelphia: Linguistic Data Consortium.
Granger, Sylviane, Estelle Dagneaux, and Fanny Meunier (eds.). 2002. *The International Corpus of Learner English. Handbook and CD-ROM*. Louvain-la-Neuve: Presses Universitaires

de Louvain.
Gries, Stefan Th. 2006. Corpus-based methods and cognitive semantics: the many meanings of *to run*. In Stefan Th. Gries & Anatol Stefanowitsch (eds.), *Corpora in cognitive linguistics: corpus-based approaches to syntax and lexis*, 57-99. Berlin & New York: Mouton de Gruyter.
Gries, Stefan Th. 2013. Data in Construction Grammar. In Graham Trousdale & Thomas Hoffmann (eds.), *The Oxford Handbook of Construction Grammar*, 93-108. Oxford: Oxford University Press.
Hanke, Thomas. 2004. HamNoSys - representing sign language data in language resources and language processing contexts. In Oliver Streiter & Chiara Chiara (eds).,:LREC 2004, Workshop proceedings : Representation and processing of sign languages. Paris : ELRA, 2004, 1-6.
Hirschmann, Lynette & Nancy A. Chinchor. 1997. MUC-7 Coreference Task Definition, version 3.0. <http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/co_task.html>.
Hunston, Susan. 2002. *Corpora in applied linguistics*. Cambridge: Cambridge University Press.
Ide, Nancy. 1998. Corpus encoding standard: SGML guidelines for encoding linguistic corpora. *Proceedings of LREC 1998*, 463-470.
Iruskieta, Mikel, Arantza Diaz de Ilarraza, & Mikel Lersundi. to appear. Establishing criteria for RST-based discourse segmentation and annotation for texts in Basque. *Corpus Linguistics and Linguistic Theory*.
Jefferson, Gail. 1978. Sequential aspects of storytelling in conversation. In Jim Schenkein (ed.), *Studies in the organization of conversational interaction*, 219-248. New York: Academic Press.
Jefferson, Gail. 1983a. Issues in the transcription of naturally-occurring talk: Caricature versus capturing pronunciation particulars. *Tilburg Papers in Language and Literature* 34.
Jefferson, Gail. 1983b. An exercise in the transcription and analysis of of laughter. *Tilburg Papers in Language and Literature* 35.
Jefferson, Gail. 1985. An exercise in the transcription and analysis of laughter. In Teun A. van Dijk (ed.), *Handbook of discourse analysis*, vol. III, 25-34. New York: Academic Press.
Jefferson, Gail. 1996. A case of transcriptional stereotyping. *Journal of Pragmatics* 26(2). 159-170.
Johnston, Trevor. 2013. Auslan Corpus annotation guidelines. <http://www.auslan.org.au/about/annotations/>.
Jorgensen, Julia. 1990. The psychological reality of word senses *Journal of Psycholinguistic Research* 19(3). 167-190.
Jun, Sun-Ah (ed). 2005. *Prosodic typology: The phonology of intonation and phrasing.* Oxford: Oxford University Press.
Kendon, Adam. 2004. *Gesture: Visible action as utterance.* Cambridge: Cambridge University Press.
Kilgarriff, Adam. 1997. I don't believe in word senses. *Computers and the Humanities* 31(2). 91-113.
Kipp, Michael, Michael Neff & Irene Albrecht. 2007. An annotation scheme for conversational gesture: How to economically capture timing and form. *Language Resources and Evaluation* 41(3/4). 325-339.
Koehn, Philipp. 2005. Europarl: a parallel corpus for statistical machine translation. MT Summit 2005.
Lücking, Andy, Kirsten Bergman, Florian Hahn, Stefan Kopp & Hannes Rieser. 2010. The

Bielefeld Speech and Gesture Alignment Corpus (SaGA). *Proceedings of LREC 2010 workshop: Multimodal corpora-Advances in capturing, coding and analyzing multimodality,* 92-98.

Leech, Geoffrey, Tony McEnery, & Martin Wynne. 1997. Further levels of annotation. In Roger Garside, Geoffrey Leech, & Tony McEnery (eds.), *Corpus annotation: linguistic information from computer text corpora*, 85-101. London & New York: Longman.

Leech, Geoffrey. 2005. Adding linguistic annotation. In Martin Wynne (ed.), *Developing linguistic corpora: a guide to good practice*, 17-29. Oxford: Oxbow.

Lu, Hui-Chuan. 2010. An annotated Taiwanese Learners' Corpus of Spanish, CATE. *Corpus Linguistics and Linguistic Theory* 6(2). 297-300.

Lüdeling, Anke & Merja Kytö (eds.). 2008. *Corpus linguistics: an international handbook. Vol. 1*. Berlin & New York: Walter de Gruyter.

MacWhinney, Brian. 2011. The expanding horizons of corpus analysis. In John Newman, R. Harald Baayen, & Sally Rice (eds.), *Corpus-based studies in language use, language learning, and language documentation*, 178-212. Amsterdam: Rodopi.

Marcus, Mitchell P., Beatrice Santorini, & Mary Ann Marcinkiewicz 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19(2). 313-330.

Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands. 2014. EUDICO Linguistic Annotator (ELAN). <http://tla.mpi.nl/tools/tla-tools/elan/>

McEnery, Tony & Nick Ostler. 2000. A new agenda for corpus linguistics – working with all of the world's languages. *Literary and Linguistic Computing* 15(4). 403-419.

McEnery, Tony, Riachard Xiao, & Yukio Tono. 2006. *Corpus-based language studies: an advanced resource book*. London & New York: Routledge.

Mitkov, Ruslan. 2008. Corpora for anaphora nad coreference resolution. In Anke Lüdeling & Merja Kytö (eds.). *Corpus linguistics: an international handbook. Vol. 1*, 579-598. Berlin & New York: Walter de Gruyter.

Müller, Cornelia. 1998. *Redebegleitende Gesten: Kulturgeschichte – Theorie – Sprachvergleich, volume 1 of Körper – Kultur – Kommunikation.* Berlin Verlag: Berlin.

Nelson, Gerald, Sean Wallis, & Bas Aarts. 2002. *Exploring natural language: Working with the British Component of the International Corpus of English*. Amsterdam & Philadelphia: John Benjamins.

Oostdijk, Nelleke & Lui Boves. 2008. Preprocessing speech corpora. In Anke Lüdeling & Merja Kytö (eds.). *Corpus linguistics: an international handbook. Vol. 1*, 642-663. Berlin & New York: Walter de Gruyter.

Ostler, Nicholas. 2008. Corpora of less studies languages. In Anke Lüdeling & Merja Kytö (eds.). 2008. *Corpus linguistics: an international handbook. Vol. 1*, 457-483. Berlin & New York: Walter de Gruyter.

Palmer, Martha, Daniel Gildea, & Paul Kingsbury. 2005. The Proposition Bank: an annotated corpus of semantic roles. *Computational Linguistics* 31(1). 71-105.

Pellard, Thomas. 2010. Ōgami (Miyako Ryukyuan). In Michinori Shimoji & Thomas Pellard (eds.), *An introduction to Ryukyuan languages*, 113-166. Tokyo: Research Institute for Languages and Cultures of Asia and Africa.

Pierrehumbert, Janet. 1980. *The phonology and phonetics of English intonation*. Unpublished Ph.D. dissertation, MIT.

Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, & Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008).

Pustejovsky, James. et al. 2003. The TIMEBANK Corpus. *Proceedings of Corpus Linguistics 2003*, 647-656.

Rayson, Paul & Mark Stevenson. 2008. Sense and semantic tagging. In Anke Lüdeling & Merja Kytö (eds.). 2008. *Corpus linguistics: an international handbook. Vol. 1*, 564-579. Berlin & New York: Walter de Gruyter.

Rice, Keren. 2012. Ethical issues in linguistic fieldwork. In Nicholas Thieberger (ed.), *Oxford handbook of linguistic fieldwork,* 407-429. Oxford: Oxford University Press.

van Rooy, Bertus & Lande Schäfer 2002. The effect of learner errors on POS tag errors during automatic POS tagging. *Southern African Linguistics and Applied Language Studies* 20(4). 325-335.

Rühlemann, Christoph & Matthew Brook O'Donnell. to appear. Introducing a corpus of conversational stories: construction and annotation of the *Narrative Corpus* and interim results. *Corpus Linguistics and Linguistic Theory*.

Sacks, Harvey, Emanuel A. Schegloff, & Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language* 50(4). 696-735.

Santorini, Beatrice. 1990. Part-of-Speech Tagging Guidelines for the Penn Treebank Project. 3rd revision, 2nd printing. <ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz>.

Schegloff, Emanuel A. 2007. *Sequence organization in interaction.* Cambridge: Cambridge University Press.

Schmid, Helmut. 2008. Tokenizing and part-of-speech tagging. In Anke Lüdeling & Merja Kytö (eds.). *Corpus linguistics: an international handbook. Vol. 1*, 527-551. Berlin & New York: Walter de Gruyter.

Slotjes, Han & Peter Wittenburg. 2008. Annotation by category – ELAN and ISO DCR. In: Proceedings of LREC 2008.

Streeck, Juergen. 2008. Depicting by gesture. *Gesture* 8(3). 285-301.

Tagliamonte, Sali. 2007. Representing real language: consistency, trade-offs, and thinking ahead! In Joan C. Beal, Karen P. Corrigan, & Hermann L. Moisl (eds.), *Creating and digitizing language corpora. Vol. 1: Synchronic databases*, 205-240. Houndmills: Palgrave Macmillan.

Taylor, Ann, Mitchell P. Marcus, & Beatrice Santorini. 2003. The Penn Treebank: An overview. *Text, Speech and Language Technology* 20. 5-22.

The British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium.: <http://www.natcorp.ox.ac.uk/>.

Thieberger, Nicholas & Andrea L. Berez. 2012. Linguistic data management. In Nicholas Thieberger (ed.), *Oxford handbook of linguistic fieldwork,* 90-118. Oxford: Oxford University Press.

Thompson, Herny S. & David McKelvie. 1997. Hyperlink semantics for standoff markup of read-only documents. *Proceedings of SGML Europe*. <http://www.ltg.ed.ac.uk/~ht/sgmleu97.html>.

University of Hamburg 2014. iLex – a tool for sign language lexicography and corpus analysis. <http://www.sign-lang.uni-hamburg.de/ilex/>.

Woodbury, Anthony. 2011. Language documentation. In Peter K. Austin & Julia Sallabank (eds.), *The Cambridge handbook of endangered languages,* 159-186. Cambridge; Cambridge University Press.

Xiao, Richard. 2008. Theory-driven corpus research: using corpora to inform aspect theory. In Anke Lüdeling & Merja Kytö (eds.). *Corpus linguistics: an international handbook. Vol. 2*, 987-1008. Berlin & New York: Walter de Gruyter.

Zinsmeister, Heike, Erhard Hinrichs, Sandra Kübler, & Andreas Witt. 2008. Linguistically annotated corpora: quality assurance, reusability and sustainability. In Anke Lüdeling & Merja Kytö (eds.). *Corpus linguistics: an international handbook. Vol. 1*, 759-776. Berlin & New York: Walter de Gruyter.