The expanding horizons of corpus analysis

Brian MacWhinney

Carnegie Mellon University

Abstract

By including a focus on multimedia interactions linked to transcripts, corpus linguistics can vastly expand its horizons. This expansion will rely on two continuing developments. First, we need to develop easily used methods for each of the ten analytic methods we have examined, including lexical analyses, QDA (qualitative data analysis), automatic tagging, language profiles, group comparisons, change scores, error analysis, feedback studies, conversation analysis, and modeling. Second, we need to work together to construct a unified database for language studies and related sciences. This database must be grounded on the principles of open access, data-sharing, interoperability, and integrated structure. It must provide powerful tools for searching, multilinguality, and multimedia analysis. If we can build this infrastructure, we will be able to explore more deeply the key questions underlying the structure and functioning of language, as it emerges from the intermeshing of processes operative on eight major timeframes.

1. Introduction

Corpus linguistics has benefitted greatly from continuing advances in computer and Internet hardware and software. These advances have made it possible to develop facilities such as BNCweb (bncweb.lancs.ac.uk), LDC (Linguistic Data Consortium) online, the American National Corpus (americannationalcorpus. org), TalkBank (talkbank.org), and CHILDES (childes.psy.cmu.edu). In earlier periods, these corpora were limited to written and transcribed materials. However, most newer corpora now include transcripts linked to either audio or video recordings. The development of this newer corpus methodology is facilitated by technology which makes it easy to produce high-quality video recordings of face-to-face interactions. Also, we now have common access to software that can link transcripts to the media at the level of the utterance and sometimes the word. These linked transcripts can then be accessed over the web, either through downloading or through browser plug-ins.

Among the various available resources, the collections at TalkBank and CHILDES are unique in providing users with completely open, online access to a large quantity of audio or video that has been linked to transcripts on the level of individual utterances. Using these links, researchers can play back thousands of different conservational interactions and study in detail their interactional and linguistic features. These materials can then be subjected to the analytic methods of corpus-based linguistics. By applying both the traditional methods of corpus

analysis and new methods specifically designed for multimedia data, we can achieve a rapid and powerful expansion of the horizons of corpus linguistics.

In this paper, I will argue that this linkage of transcripts to media opens up a unique, new way of understanding human language. My basic thesis is that language emerges through the interaction of forces that operate on several different time scales or timeframes. These timeframes vary in their duration from milliseconds to millennia. However, all of these forces must eventually reveal their functioning within the actual moment of communication. Collections of transcripts linked to video provide us with a methodologically sound way of tracking and evaluating these interactions between timeframes across situations, speakers, and time. In this way, the availability of transcripts linked to video can radically expand the horizons of corpus linguistics. However, to realize these expanded horizons, researchers must commit themselves to the construction of a shared, interoperable database, similar to that currently available in TalkBank. Here, I will explain how this database must be structured to best address the many research challenges we will face within this broadened scope of corpus linguistics.

2. Timeframes meshing in the moment

Let me begin by posing the core question facing linguistic analysis. In his studies of sentence production, Osgood (1971) phrased the question as "Where do sentences come from?" In a similar vein, we can ask, "Where do linguistic forms and functions come from?" A full answer to this question would provide a detailed functionalist account of the shape of human language. In this paper, I will suggest that functions and forms arise from the competing intersection of adaptive processes operating across interlocking timeframes. We can distinguish eight major levels of timeframes; each of these levels includes a variety of individual mechanisms whose timeframes also vary:

1. **Processing**. This timeframe governs the online processing of words and sentences.
2. **Turn-taking**. This timeframe governs how we maintain, complete, extend, and yield conversational turns.
3. **Activity**. This timeframe governs how we structure the overall goals of the current interaction.
4. **Developmental**. This is the timeframe governing language learning by the child and the adult second language learner.
5. **Social**. This is the timeframe within which we track commitments to various social groups and processes across interactions.
6. **Epigenetic**. This is the timeframe across which the genes interact to produce physical and neural structures supporting language.
7. **Diachronic**. This is the timeframe for socially shared changes in language forms and structures.

8. **Phylogenetic**. This is the timeframe that governs the ongoing evolution of the human language capacity.

The shapes of specific linguistic markers are primarily determined by forces on levels 1-3, but are further modulated by the long-scale processes on timeframes 4-7. All of these functions receive inputs from all timeframes. For example, the choice between a noun, pronoun, or zero in subject position is conditioned by forces for the sharing of mental models operating on timeframes 1, 2, and 3. These forces are constrained at the moment of speaking by grammatical and production processes within timeframe 1. Across the timeframe of language development, the growth of control over pronominal marking arises within timeframe 4 and changes in accord with timeframe 5. Finally, the underlying function of deixis and pointing has emerged in the species across timeframe 8 with ongoing support from all the other levels.

A key postulate of functionalism (Bates & MacWhinney 1989) is that functions compete for mapping to forms. In the timeframe model, the outcome of the competition is determined by inputs from clocks running at each time scale. The brain provides support for this type of learning by using circuits with widely divergent time scale properties operating between the basal ganglia, hippocampus, and the cortex. Even in organisms as simple as the bee, we can find discrete neurochemical processes that work to consolidate memories on each of the time scales relevant to the pollen-gathering activities of the bee (Menzel & Giurfa 2001).

For the purposes of illustration, one can compare the timeframes involved in control of language with those used in mechanical calendars. An example from ancient Greece is the Antikythera, a device discovered in 1901 inside a shipwreck in the Aegean. The device uses a set of interlocking gears or wheels to compute times according to the 365-day Sothic year, the 12-month zodiac, the solar year, the lunar phase, the phases between solar eclipses, the 19-year Metonic cycle, and the 76-year Callippic cycle. In addition, it marks the positions of the planets and the times of specific Panhellenic games. Figure 1 provides a glimpse into the shape of this remarkable mechanism. Another example of a clock that meshes across multiple time scales is the Orloj from Prague, which is shown in Figure 2.

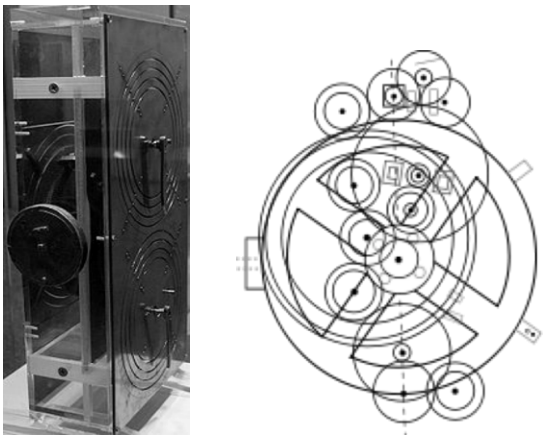

Figure 1: The Antikythera: a reconstruction and a sketch of the gears
Images accessed at Wikipedia Commons:
http://en.wikipedia.org/wiki/Antikythera_Mechanism

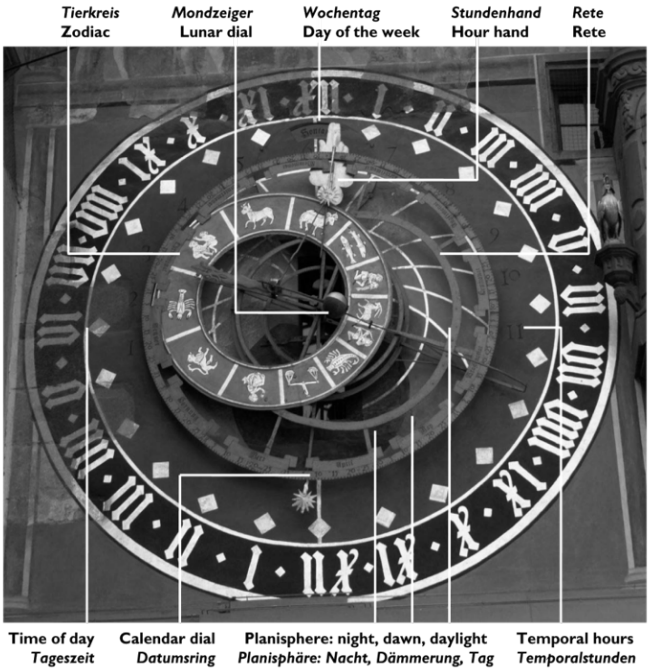

Figure 2: The Orloj clock of Prague
Image accessed at Wikipedia Commons:
http://en.wikipedia.org/wiki/Prague_Astronomical_Clock

The meshing of the wheels of these mechanical calendars provides an analogy to the meshing of the processes operating across the eight time scales determining human communication. The notion of process meshing can also be illustrated through biological systems. For example, metabolic processes in the body link up across a variety of time scales, some tightly linked to the catalytic processes in the Krebs cycle that convert ATP into energy, others involved in the creation of ATP, and still others involved in the overall regulation of the body's heat and energy levels. Even more complex and longer-term regulatory processes govern learning in the brain and its regulation by hormonal systems and exposure to inputs from the environment.

Let us now take a closer look at the shapes of these eight timeframes with attention to how they affect particular linguistic structures or processes. Within each of these timeframes, when we look at specific examples, we will see that there is variation in the actual clocks involved. This is why it is important to use the term "timeframe" to discuss a general class of processes that operate with similar, but not identical, clocks.

2.1 The processing timeframe

The most fast-acting pressures on language form are those that derive from online processing constraints (MacWhinney 2009). These pressures arise from the limitations of memory mechanisms, attention focusing, coordination of sentence planning, code switching between languages, and motor control. These various processing timeframes all rest on top of the basic time required for the firing of a single neuron and transmission of that signal to the next neuron. Although this depends heavily on many factors, a value of 10 milliseconds (ms) is a good approximation for many connections. The brevity of this period allows for the firing of as many as 20 neural connections during the 150 ms period required for production of a single syllable (Massaro 1987). Consider this example: when bilinguals switch from English to Spanish, the initial 250 ms of speaking in Spanish are still under the influence of English phonology until the alternative Spanish patterns become fully active (Grosjean & Miller 1994). Similarly, Goldman-Eisler (1968) found that the majority of retraces and repetitions typically affect nothing more than one or two syllables, extending across perhaps 500 ms. Attentional shifting between perceptual dimensions or languages can also be measured in terms of hundreds of milliseconds (Prior & MacWhinney in press).

The processing timeframe brackets a large number of important processes in language production and perception. In terms of production, this timeframe governs lexical access (including gang effects and competitor effects), phonological activation in the output buffer, morphological combination, and triggering of syntactic combinatorial patterns. This timeframe also governs phonological assimilations occurring in fast speech patterns (Bybee 2003). Using corpus linguistic methods, we can study how these assimilations increase through a discourse and how they vary with the nature of the addressee. In this way we can see interactions between timeframes 1, 2, 3, and 5. When we start to look in

detail at fast speech assimilations in the speech mothers provide to their children, we then see further interactions with timeframe 4 (Smith, Durham & Fortune 2007).

On the perceptual side, the processing timeframe brackets all of the basic auditory processes, as well as the input to statistical pattern learning, short-term auditory memory, processes for segmenting by stress patterns, orientation to specially marked signals, and adjustment to speaker variability. Generally, it is easier to apply corpus-based methods to the study of production. However, we can also study perception by using corpora to delineate the shape of the ambient language that serves as input to the learner and the framework for social variation.

2.2 The turn-taking timeframe

When we are engaged in conversational interactions, we are continually taking in data from our conversational partners. We are watching their postures, gestures, and facial expressions. As we are speaking, they may be inserting nods or words indicating their acceptance of what we are saying or at least showing that they are paying attention. However, they may also be indicating impatience or a desire to have the floor. At each point in the conversation, we are involved in the construction and completion of our current turn constructional unit (TCU) (Schegloff 2007). Through syntactic structure, gesture, and prosody we can signal our desire to maintain or yield the floor. When a TCU is completed, we may add a further increment, start a new TCU, or allow an interlocutor to begin speaking. All of these decisions involve the continual ingestion of information by the turn-taking system during real time. The timeframe here is a bit longer than that of basic linguistic processing. The basic system that responds to feedback from listeners operates in a framework of 500 ms, with additional components extending across a few seconds. By studying video data from natural, spontaneous interactions, corpus linguistics can make serious contributions to the understanding of processes in this timeframe. However, achieving this progress requires adherence to a set of coding conventions for conversational features that can be unambiguously processed by computers. This goal has been addressed within the CABank component of TalkBank at http://talkbank.org/CABank.

2.3 The activity timeframe

Beyond our short-term involvement with turn-taking, we also make medium-term commitments to the shape of an interaction as an activity. For each interaction, we set conversational goals that define particular activity types (Leont'ev 1947/1981). For example, we may engage a real estate agent to help us buy a house. Our linguistic interactions with this agent are then shaped by the status of the buying process and our goals in buying a house. After we complete one set of transactions with this agent, we will maintain an ongoing relation that will then shape our further interactions, days or weeks later (Keenan, MacWhinney & Mayhew 1977). Activity types provide a major challenge and opportunity for

corpus-based analysis, because the shape of the linguistic resources we use vary widely across activity or situation types. For example, the default reading of the word "bank" will be different if we are discussing economics, as opposed to methods of improving river hydraulics. If we could configure computational corpus analysis to be sensitive to alternative activity types or genres (Biber, Conrad & Reppen 1994), we could markedly improve speech recognition, ambiguity resolution, parsing, and automatic discourse processing.

2.4 The developmental timeframe

Jean Piaget's genetic psychology (Piaget 1954) was the first fully articulated emergentist view of development. Impressively complete in its coverage, it failed to specify details regarding mechanisms of development. To provide this missing mechanistic detail, current emergentist accounts of development rely on connectionism (Quinlan 2003), embodied cognition (Klatzky, MacWhinney & Behrmann 2008), and dynamic systems theory (Thelen & Smith 1994). Emergentist theory has been used to characterize two different, but interrelated, aspects of development. The first is the learning of basic facts, forms, relations, names, and procedures. Connectionist and usage-based models of language learning, such as those that deal with learning of the past tense (MacWhinney & Leinbach 1991), syntactic patterns (Waterfall, Sandbank, Onnis & Edelman 2010), or word segmentation (Monaghan & Christiansen 2010) often focus on this type of development. A second type of development involves the learning of new strategies and frameworks that can alter the overall shape of language and cognition, often through cue focusing and bootstrapping (Regier 2005, Smith & Colunga 2003). The relevance of corpus-based analyses to the study of language development has been demonstrated extensively in recent work based on the CHILDES and BilingBank corpora. There are now over 3,500 published articles based on the use of these corpora. A recent issue of the *Journal of Child Language* (MacWhinney 2010) based on the modeling of corpus data from CHILDES illustrates the high level of analytic precision that this work is now achieving.

The developmental timeframe meshes in complex ways with processing during the ongoing interaction. For example, a child's failure to understand the meaning of the word "dependability" in a classroom discussion of the reliability of batteries (MacWhinney 2005) may be the result of problems in understanding previous classroom and computerized lessons on numerical distributions. Similarly, the failure in lexical retrieval that occurs in aphasia is driven by changes to neural tissue subsequent to a stroke. Thus, online processing emergence can reflect the status of long-term developmental, neuronal, and physiological processes.

2.5 The social timeframe

Many of the pressures that operate during face-to-face conversations derive from long-term social commitments. Our choice of vocabulary, slang, topics, dialect

and language is determined by the status of our social relations to the people we meet. We can select particular linguistic options to emphasize solidarity (R. Brown & Gilman 1960), impose our power, or seek favors (P. Brown & Levinson 1987). The time course of these social commitments is often measured in years or decades (Labov 2001). Some basic social commitments, including those forced by gender and race, can never be fundamentally altered. For corpus-based analyses, what is important is being able to link the usage of linguistic forms in the moment to long-term personal characteristics, whether observed directly or measured through questionnaires and surveys. This is the approach taken within the field of sociophonetics (Hay & Drager 2007) which studies the meshing of long-term social indicators with output processing in the moment. To make corpora maximally useful for such studies, we need both good tools for phonetic analysis and systematic coding of the relevant social variables. To achieve this we have been working to configure tools such as Praat (praat.uva.nl), Phon (talkbank.org/phon), and CLAN (talkbank.org/clan) to integrate the phonetic and sociolinguistic aspects of these analyses.

2.6 The epigenetic timeframe

DNA lays out framework that governs the epigenesis (Waddington 1957) of the embryo and the infant. However, this is an emergent framework, rather than a detailed blueprint. Embryologists have shown how biological structures emerge from processes of induction between developing tissue structures in the embryo in cascading waves of catalytic processes (Goodenough & Deacon 2006). The shape of these interactions is not hard-coded in the DNA. Instead, the DNA encodes information that can push the process of differentiation in particular directions at crucial epigenetic choice points. The precursors of autism in the embryo can be traced to particular epigenetic effects, as can the formation of stripes in the tiger. Epigenetic emergence does not cease at birth. To the degree that the brain maintains a level of plasticity, epigenetic processes allow for recovery of function after stroke through rewiring and reorganization. Before birth, epigenetic interactions with the environment are confined to forces that impinge on the uterus and the embryonic fluid. After birth, the environment can trigger a wide variety of variations in gene expression from diabetes to brain reorganization for language in the deaf (Bellugi, Poizner & Klima 1989). As long as the child remains in the womb, few of these processes are relevant to work in linguistics. However, once the infant begins vocalizations and babbling, there is a great deal that can be studied through corpus-based methods, as recently codified in the context of the Phon Project (talkbank.org/phon) organized by Yvan Rose at Memorial University.

2.7 The diachronic timeframe

We can also use emergentist thinking to understand the changes that languages have undergone across the centuries (Bybee & Hopper 2001). Typically, these changes are viewed as resulting from the fast speech processing simplifications or

morphological overregularizations. It is the cumulative effect of these myriad short-term effects that drive the long-term changes. However, the distribution of these processes is not necessarily uniform. For example, Labov (2001) notes that teenage males are particularly involved in the generation of new speech patterns. To understand this better, we need to collect corpora that capture the variation in the speech forms produced by these highly active groups.

2.8 The phylogenetic timeframe

The slowest moving emergent structures are those that are encoded in the genes. Changes across this timeframe – which involves millennia rather than minutes – are controlled by natural selection (Darwin 1871). The core engine of emergence is the generation of variation through mutation, followed then by natural selection through both mate choice and differential mortality. Natural selection utilizes the possibilities for reorganization shaped by the DNA and the interactions of polypeptides that it specifies. The unevenness of this underlying landscape makes some mutations more probable and frequent than others, leading to a reliance on the reuse of old forms to serve new functions. Emergentist accounts in this area have emphasized the ways in which language, society, and cognition have undergone coevolution (MacWhinney 2008a) based on the linking of dynamic systems. To trigger this coevolutionary advantage, changes in linguistic abilities must arise in parallel with advances in cognitive or social abilities. Moreover, both effects must interact at the moment of speaking. When this happens in a way that favors reproductive fitness, the mutation will be preserved. To study the linkage of this timeframe to interaction in the moment, researchers must focus on individual and population differences. For instance, one can study clinical patterns such as Specific Language Impairment (SLI), stuttering, autism, or Williams Syndrome in terms of their impact on communication during the moment. As an example, there is evidence that Specific Language Impairment arises from difficulties in integrating diverse types of linguistic information online (Presson & MacWhinney in press). One important line of research seeks to link these patterns to specific genetic patterns in genes such as FOXP2 (Enard et al. 2002). However, we also need to trace the effects of these genetic factors on language both through controlled experimental tasks and corpus analysis. Corpus analysis allows us to track ways in which the language of these populations differs from the standard pattern. By examining these patterns in detail, we can learn which constructions trigger errors, avoidance, or alternative strategies.

2.9 Linking the timeframes in the moment

All of these processes, across all of these timeframes, must have their impact at the moment of speaking. This means that if we can capture the moments of speaking in sufficient quality and quantity, we will be able to study the interplay of all of these processes. To achieve this, the primary TalkBank data structure is an alignment between a digitized transcript (or annotation) and digitized media, much as in the TalkBank database. As Bird & Liberman (2001) have shown,

linkages of annotations to media all assume the form of directed acyclic graphs (DAGs) which they call "annotation graphs" (AG). Although this structure is universal, there are dozens of methods for characterizing and displaying these transcript-media linkages. The next page shows two ways in which a transcript can be linked to media. In the first format, from ELAN (www.lat-mpi.eu/tools/elan), we see the transcript displayed in a musical notation format under the QuickTime window (Figure 3). In the second format, from CLAN (childes.psy.cmu.edu/clan), we see the transcript in a standard textual display (Figure 4). In both cases, as the researcher plays through the video, the respective segment of the transcript is highlighted so that the researcher can study the transcript in tight synchrony with the actual experience of the interaction.

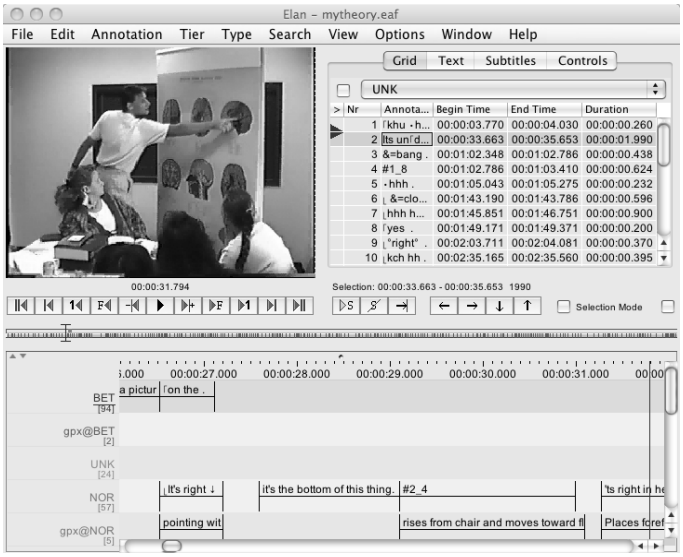

Figure 3: A time-aligned transcript as displayed by ELAN

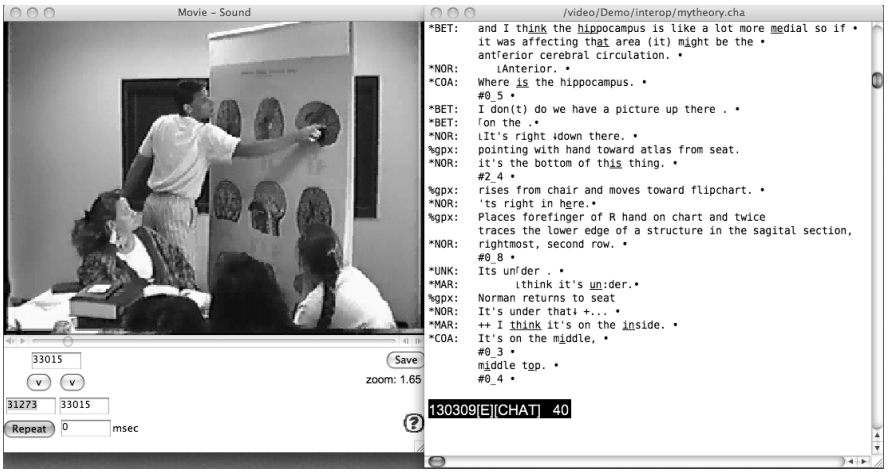

Figure 4: A time-aligned transcript as displayed in CLAN

This particular videotaped interaction was the subject of a special issue of *Discourse Processes* in 1999. In the interaction, medical students in a problem-based learning (PBL) class are trying to link the position of the hippocampus to observed symptoms in a neurological patient. The CD-ROM at the back of the special issue presented the transcripts linked to the media. These transcripts with linked media can be browsed over the web from http://talkbank.org/browser. This particular example is at ClassBank/CogInst/mytheory.cha.

The advantage of the ELAN musical score display format is that it displays clearly the duration of a communicative activity and the synchrony across types of communicative activities (intonation, head nodding, words, etc.). The advantage of the CLAN transcript format is that it is easier to read and presents a better overview of the conversation. Each visualization format has its own advantages and disadvantages. The important point is that, with full data interoperability, researchers can display the same underlying data in whichever viewer or editor they need for the particular project at hand. Toward this goal, we have written programs that can reliably convert data from all of the major transcript-media display programs (AG, Anvil, ELAN, EXMARaLDA, HIAT, SALT, DRT, SyncWriter, MediaTagger, Transcriber, Transana, Praat, WaveSurfer, SoundWalker, and Observer) into the ComNet XML format. Moreover, transcriptions and other annotations can be transformed from XML back into each format and compared with the original to show that no data were lost during the roundtrip conversion.

3. Analysis methods

Corpora composed of multimedia interactions linked to transcripts can be analyzed through a wide variety of methods. Because only some of these methods involve controlled experimentation, there is sometimes a tendency to dismiss corpus analysis as unscientific or pre-scientific. However, many of the most important advances in science have come from non-experimental methods, such as naturalistic observation (Darwin, Linnaeus), model-building (Hawkins), and thought experiments (Einstein). Moreover, corpus analysis can easily be combined with experimental control, when necessary. Therefore, any general attempt to dismiss corpus analysis as unscientific is inappropriate. Instead, methodological critiques should focus on understanding the match between particular methods and their individual scientific goals. In this section, I will review 12 classes of methods that have proven useful in the study of multimedia corpora. To illustrate these methods, I will often refer to specific computational implementations available in the CLAN programs (talkbank.org/CLAN) created by the TalkBank project. In fact, there are dozens of other computational implementations that overlap with features of the CLAN programs and I am using CLAN here only to provide a consistent source of examples.

3.1 Lexical studies

The most widely used methods in corpus analysis involve the study of the distribution of lexical forms. What makes these methods particularly appealing is the ease with which computers can locate specific surface word forms in corpora. Internet search engines such as Google have taken advantage of this fact to construct indices to the vast textual resources of the Internet. These same methods can also be applied to multimedia resources with transcripts linked to video. The various lexical methods fall into several categories or levels, based on the extent to which they rely on additional structures:

1. The simplest lexical methods treat corpora as mere "bags of words". From this bag of words, we can compute lists of word frequencies, based simply on surface forms. For example, the FREQ program in CLAN provides frequency lists that can be ordered alphabetically or by frequency. As in BNCWeb, the output of FREQ can be used to locate the usages in the original transcripts. Moreover, the output can be shipped over to programs such as Excel for further statistical analysis.
2. Beyond the first level of lexical counting, it is easy to construct a specific search set that targets words from a given semantic domain. For example, FREQ allows the user to track the frequencies of words in specific user-defined lexical groups, such as "morality terms" or "quantifiers".
3. One can also characterize texts in terms of a variety of lexical profiles. For example, the type/token ratio (TTR), which is also produced by the FREQ program, can be used to indicate the lexical diversity of a text.

However, this index is not constant across sample sizes, because the impact of high frequency words on lowering the TTR is most clearly demonstrated in larger texts. To correct for this, CLAN encourages the user to rely instead on the VOCD measure of vocabulary diversity (Malvern, Richards, Chipere & Purán 2004).

4. On the next higher level, CLAN provides methods of searching lexical combinations through regular expression (Regex) pattern matching (Friedl 2002). Using the KWAL and COMBO programs, users can locate each match to a pattern and double click on the entry to go to relevant line in the original transcript. At that point, they can study the discourse context and play back the audio or video, if needed.
5. In many cases, it is possible to formulate research questions in terms of the surface forms of words. However, in other cases, researchers want to study lexical combinations in relation to part of speech (POS) tags and syntactic tags. To support this, many of the corpora in CHILDES have been tagged for both morphosyntactic composition and syntactic role. These tags are aligned with words on the main orthographic tier in a one-to-one fashion, as verified through the XML validator. The online WebBNC corpus (bncweb.info) provides a good example of how simple Regex searches can be combined directly with POS information. The current versions of COMBO and KWAL in CLAN can perform some of these searches. However, the newer CHATTER program provides fuller support for XML-based Regex queries. Unlike CLAN, which is written in C, CHATTER is written in Java and designed specifically for searching through the XML version of the various TalkBank databases.

These various methods for lexical searching all depend on the use of consistent standards for lexical representation. For spoken language corpora, it is often difficult for transcribers to adhere to a consistent set of standards for lexical representation. MacWhinney (2008b) describes these problems in detail. Many of them can be resolved through simple standardization. However, the most serious problem relates to the desire to represent the actual phonological form of the word directly in the lexical line. Because TalkBank transcripts are typically linked to audio, the need for representing phonology directly through eye-dialect (Ochs 1979) is minimized. In addition, CHAT provides methods for representing phonological form, but then following this with the standard lexical target. Because of these various problems, many of the corpora in the CA segment of TalkBank may never be subjected to POS tagging. Eventually, however, the bulk of the corpora in CHILDES and other segments of TalkBank will be tagged.

3.2 Qualitative Data Analysis = hand coding

Corpora can also be subjected to analysis through coding systems. Areas that rely heavily on the use of hand coding include studies of gesture, speech acts, dialect features, interactional patterns, rhetorical structure, and speech errors. The process of elaborating transcripts with hand-entered codes is known as

Qualitative Data Analysis (QDA). Major programs supporting QDA include NVivo7, ATLAS.ti, Kwalitan, The Observer, HyperResearch, and Transana. Information on these programs is given at http://talkbank.org/software/. Unfortunately, these programs all rely on proprietary transcript formats, and are not linked to any publicly accessible database. As a result, researchers who are interested in coding corpora currently in TalkBank need to rely on the QDA tools provided in CLAN. These tools include the following components:

1. The Coder's Editor program facilitates exhaustive coding of a transcript in accord with a user-specific system. The researcher first specifies a hierarchical coding structure. A good example of this is the INCA speech act coding system (Ninio & Wheeler 1984) given in the CHAT manual. Once a coding system has been specified, the coder then goes through the transcript, line-by-line making a selection of the correct code for each utterance.
2. After using Coder's Editor, the RELY program allows one to compute reliability across two or more coders.
3. GEMs are marks the coder inserts in the transcript to surround interesting blocks of text with code words such as "book reading" or "political discussion". Later on, programs can use these marks to focus analyses on the relevant blocks or "gems".
4. CLAN also allows users to insert codes on any number of additional coding lines such as %coh for cohesion analysis, etc.
5. The TalkBank system also provides a method for work groups to enter blogs or commentary for multimedia-linked transcripts on the web at http://talkbank.org/browser. After registering for this system (MacWhinney et al. 2004), users can see comments from others in their work group and add additional comments of their own. Commentary can also be coded by type, date, and other parameters.
6. Gesture analysis in systems such as ELAN (http://www.lat-mpi.eu/tools/elan/) or Anvil (www.anvil-software.de) provides very accurate alignment of speech to gesture. To interface with these programs, CLAN provides converters between CHAT XML and ELAN or Anvil XML. There are good reasons to consider use of both CLAN and ELAN (or Anvil) for gesture analysis. Within ELAN and Anvil, the screen territory tends to become cluttered with coding tiers. Moreover, it is difficult to get a standard textual overview of the interaction in combination with the gestures. To address this problem, CLAN allows transcribers to nest small files within larger files using hot links. Within the overall transcript for a session, the user clicks a bullet to open up a second embedded transcript, much like zooming in with a microscope. At the second level, a gesture sequence is described and coded in detail and further bullets are available to play back segments of the gesture from thumbnail icons or to open up a third level of gesture analysis. This

Figure 5: Main text window for CLAN gesture analysis sample

Figure 6: Sub-window for CLAN gesture analysis and coding

approach tends to focus on the analysis of gesture sequences, as in Kendon (1982). The screen shots in Figures 5 and 6 show what this coding looks like for an interaction in Danish. The main text window is shown in Figure 5. When the user clicks on the line labeled "torture time sequence", the window in Figure 6 pops up allowing for detailed coding of the gesture sequence for that passage.

3.3 Automatic tagging

As corpora grow in size and as research questions proliferate, it becomes increasingly important to enter codes automatically, rather than tediously by hand. For some domains, automatic coding is still a distant goal, but for others it is possible now. The three areas for which CLAN can now provide automatic coding or tagging include morphosyntax, grammatical relations, and phonology.

1. **MOR and POST**. For morphological tagging and disambiguation, CLAN relies on the MOR and POST programs, as documented in MacWhinney (2008b). MOR provides not only a tagging of the part of speech of the word, but also a full analysis of its morphological composition. We now have open-source MOR taggers with user-modifiable lexicons and rule systems for Afrikaans, Cantonese, Dutch, English, French, German, Italian, Japanese, Mandarin, and Spanish. After tagging by MOR, the user runs the POST disambiguator (Parisse & Le Normand 2000). The accuracy of the disambiguation is well above 95%, as long as the transcription is accurate on the word and utterance level.
2. **GRASP**. After analysis by MOR and POST, it is possible to run the GRASP program (Sagae, Davis, Lavie, MacWhinney & Wintner 2010) to automatically construct a labeled dependency graph, based on a set of user-specified grammatical relations (GRs), such as Adjunct, Subject, or Complement. Although we have GRASP taggers now for English, Japanese, Spanish, and Mandarin, each of these systems is new and further work is needed to improve tagging accuracy.
3. **Phonology**. The CLAN programs have only the most limited abilities to analyze phonological features. To address this problem, we have built a separate Java program called PHON (Rose et al. 2005) that can directly analyze files in the CLAN XML format. Using IPA transcriptions, PHON can perform a variety of automatic analyses, such as segmentation, syllabification, and process analysis. In addition, it provides automatic linkages to Praat for phonetic analysis.

In the future, we hope to extend our automatic analyses to include systems for automatic video analysis (Hauptmann, Yan, Lin, Christel & Wactlar 2007), psychological mood coding (Pennebaker, Francis & Booth 2001), propositional

tagging (Palmer, Gildea, & Kingsbury, 2005), and emotional coding (Wiebe, Wilson, & Cardle 2005).

3.4 Language profiles

Using the codes created by MOR, POST, and GRASP, one can automatically compute various language profile indices that were earlier computed tediously by hand. In particular, we have automated the computation of the Developmental Sentence Score (DSS) of Lee (1966) and the Index of Productive Syntax (IPSyn) of Scarborough (1990). These two language profile measures are important for the characterization of the normal course of child language development and the diagnoses of deviations from this normal course. For the study of aphasic language types in the context of the AphasiaBank project, we have constructed the MORTABLE program that automatically extracts a complete tabulation of POS and GR tags for further analysis in Excel. These measures are used to compare individual patients to a wider set of norms found in the one-hour protocols collected from the 150 patients currently in the AphasiaBank database.

3.5 Group and interlocutor comparisons

Many TalkBank data sets were collected in the context of systematic experiments, often involving between-group comparisons. Most of the corpora in the Clinical segment of the CHILDES database include samples from both children with language disorders and normally developing children. For example, the corpora from children with language disorders contributed by Gina Conti-Ramsden (Jones and Conti-Ramsden 1997) involve a comparison between children with Specific Language Impairment (SLI) and their younger language-matched siblings. Also, the AphasiaBank corpus has been structured to highlight the basic comparison between aphasics and age-matched normal controls. In other studies, there is a comparison between children from alternative socioeconomic, racial, or ethnic groups. For example, the Hall corpus (Hall & Tirre 1979) includes children from these four groups of families: Black Professional, Black Working Class, White Professional, and White Working Class. Other corpora compare younger and older mothers, or mothers with various levels of education. For all of these group comparison studies, there is an emphasis on maintaining a consistent method of data collection to maximize comparability between the groups. In the most extreme cases, this involves use of a constant, tightly specified protocol, including specific tasks and questions.

Some of the sociolinguistic studies in TalkBank rely on comparisons between interlocutors, rather than subject groups. For example, the Gleason corpus examines children's language use in three situations: talking with their mother, talking with their father, and interacting with the whole family during dinner conversations. Within the various bilingual corpora, there is often a comparison between interlocutors from different languages or situations that require the use of one language rather than the other (Yip & Matthews 2000). Within this framework, there can also be experimental manipulations that

specifically encourage reliance on one language over the other (Nicoladis & Genesee 1998).

3.6 Change scores

The most common use of change scores is to measure the naturalistic course of language development over time. For example, language acquisition researchers use measures such as lexical diversity (VOCD), mean length of utterance (MLU), or developmental sentence score (DSS) to track the course of language development for both first and second languages. These measures can be applied to either longitudinal corpora collected across several years from individual learners or cross-sectional corpora collected at single time points from groups of learners at different ages or levels of first and second language learning. These data can then be modeled to study patterns such as the vocabulary spurt (Bates & Carnevale 1993, Ganger & Brent 2004), learning bursts (Larsen-Freeman 2006), or implicational patterns (Pienemann et al. 2005). The study of change scores over development is the single most important method in the study of both first and second language acquisition.

This basic use of change scores does not involve any clear experimental treatment. However, it is also possible to combine the study of change scores with experimental control. For example, in the context of the AphasiaBank project, researchers apply transcortical magnetic stimulation to improve lexical access in aphasia. To measure the effects of this treatment, participants retell the Cinderella story (MacWhinney et al. 2010) both before and after treatment. The two retellings are then compared using automatic methods such as VOCD and MORTABLE. Differences in the pretest and posttest retellings are also contrasted with those from a control group that did not have transcortical magnetic stimulation.

Change scores can also be used to measure the effects of educational treatments. For example, in their Fluency corpus, de Jong and colleagues examine the effects of production fluency training using the 4-3-2 method (Maurice 1983, Nation 1989). Improvement in fluency after training can be measured in terms of decreasing pause length and reduced errors and omissions.

3.7 Error analysis

Language acquisition studies often focus on ways in which learner corpora diverge from the standards of the target language. In studies of second language learning, Corder (1983) placed an emphasis on the value of Error Analysis. However, other researchers noted that learner grammars are not merely degenerate versions of the target grammar, but have their own systematicity, as expressed in the notion of a learner's interlanguage (Selinker, Swain & Dumas 1975). However, even if one views the learner's productions as structured into an interlanguage, it is helpful to compare divergences between that interlanguage and the target language. To facilitate this type of analysis, the CHAT coding system provides a hierarchical system of error coding. This system has been most

extensively and consistently applied to the corpora in AphasiaBank. Aphasic speech involves many types of phonological and grammatical errors only rarely found in speech from normal child and adult learners. However, first and second language learners also demonstrate unique patterns of errors. The goal of the TalkBank error-coding system is to be applicable to coding for each of these groups.

3.8 Feedback studies

The study of learners' errors leads directly to the study of ways in which conversational partners can provide corrective feedback as well as positive exemplars (MacWhinney 2004). The most extreme cases here arise when the learner produces an overt formal error, such as "goed" and the parent or teacher makes an overt formal correction by saying, "No, you can't say 'goed', you have to say 'went'". There is extensive debate in the literature about the extent to which such overt correction occurs and whether it has a positive impact on the learner. Apart from this form of correction, there has also been study of less explicit correction through recasting (Bohannon, MacWhinney & Snow 1990). Researchers also believe that learners can make effective use of clear patterns of language presentation in the form of "variation sets" (Waterfall et al. 2010). In these sets, parents illustrate language patterns through item and replacement schemes. Yet another interactional pattern that is thought to facilitate language learning is the process of fine-tuning that occurs between parents and children. When parents become aware that their children are making use of some new syntactic or lexical pattern, they then increase their own use of this pattern on a level that is appropriate for the child (Sokolov 1993). CLAN programs such as CHIP, CHAINS, and KEYMAP can be used to track these ways of providing fine-tuning, feedback, and examples.

3.9 Conversation Analysis

TalkBank corpora are now also being used for research within the framework of Conversation Analysis (CA) (Schegloff 2007). Beginning in 2002, the CHAT transcription method was adapted to include all of the conventions of standard Jeffersonian CA transcription (Jefferson 1984). Specific Unicode symbols were introduced to mark features such as pitch rise and fall, terminal contours, overlap alignment, and various vocal qualities. These markings, which are summarized in the table at http://talkbank.org/CABank/codes.html, can all be entered through simple pairs of keystrokes. The DK-CLARIN CA Project organized by Johannes Wagner has created "gold standard" CA transcriptions for the Danish SamtaleBank collection and has reformatted Gail Jefferson's Newport Beach and Watergate transcriptions into CA CHAT. The TalkBank CA corpora now also include dozens of other corpora from group conversations, phone calls, and classroom interactions. Like the other corpora in TalkBank, these CA transcripts are linked directly to audio or video media. This linkage allows for greater analytic power in three directions. First, it is now easier to listen repeatedly to

specific segments to detect conversationally relevant signals and patterns. This can facilitate analysis by individual researchers, discussion in data sessions, and the sharpening of collaborative commentary over the web. Second, it is now easier to link transcript analysis to phonetic analysis, by directly sending individual segments to Praat for numerical analysis. Third, with the construction of large databases of CA transcripts, it is possible to search for conversational features across interactions. However, to achieve this goal, transcripts must be adapted to display features that can be easily identified by regular expression searches.

3.10 Modeling

Data analyzed using these various methods can be further examined through computational modeling. The goals here are either to account for learning data or to account for conversational patterns. Most current models focus on trying to account for learning of either first or second languages. These models usually assume that learning involves the abstraction of patterns from the input data. This means that the first step involved in building such models is to derive a good picture of the input to the learner. For example, models that seek to account for the learning of the English past tense have first attempted to characterize the frequency of different verb forms in CHILDES corpora. The second step in building these models is to determine the developmental patterns in learners. Once these two profiles have been extracted from the corpora, the model is used to show how the learner's interlanguage can emerge from processing of the input. Neural network models have been used to account for morphological learning (MacWhinney & Leinbach 1991), segmentation (Blanchard, Heinz, & Golinkoff 2010, Monaghan & Christiansen 2010), prosodic patterns (Gupta & Touretzky 1994), lexical organization (Li, Zhao & MacWhinney 2007), and aspects of syntactic development (Elman 1993). Other models have relied on discrimination nets (Ling & Marinov 1993), pattern associators (Gillis, Daelemans & Durieux 2000), hierarchical Bayesian models (Perfors, Tenenbaum & Wonnacott 2010), and basic string comparison methods (Freudenthal, Pine & Gobet 2010). What is common across all of these applications and models is the reliance on data from corpora that include both learners' speech and the speech addressed to them.

4. Building a general database

Recently, researchers associated with the LDC (Linguistic Data Consortium) and the TalkBank Project have developed a proposal for the construction of a general database for the study of human language. This section explains the overall shape of this proposal. Although funding for this project has not yet become available, aspects of this work are progressing as separate projects. The details of the proposed work will undoubtedly change over the course of time, as we learn which ideas are more successful than others. However, in its current form, this

proposal can provide us with a vision of what it means to expand the horizons of corpus linguistics.

At the core of this proposal is the observation that, each day, we generate billions of rich communication data streams in the form of conversations, meetings, broadcasts, newspapers, scientific articles, and links to scientific data. These streams of communication can be captured digitally and enriched automatically through systems for linguistic and video annotation. Databases such as TalkBank, CHILDES, Informedia, Let's Go, and LDC have demonstrated how this can be done, amassing the world's largest collections of data on spoken and written communication. However, to properly integrate these materials into a technologically sophisticated network, we must achieve full interoperability, universal access, deep curation, powerful visualization, and wider coverage.

By linking these resources together, we can build an international database for digital data on human communication, including speech (conversations, interviews, radio broadcasts, etc.), text (historical and modern books, journals, newspapers, web pages, blogs, etc.), scientific communication (articles, letters, debates, commentary, reviews, linked to primary scientific data), video, and images. This database, called ComNet, will provide material for students, teachers, businesses, professionals, language communities, government, and the public.

4.1 Basic principles

ComNet will construct a fully interoperable set of tools for data acquisition, curation, analysis, and visualization based on the TalkBank (talkbank.org/talkbank.xsd) XML schema, which will be renamed as the ComNet XML Schema. Using these tools, we will link together an international distributed database on human communication, structured in accord with these basic principles:

1. **Open access.** The core components of ComNet will be freely open to all – students, researchers, government, private industry, and the public. For non-core components, access may be limited by LDC licensing, TEACH Act requirements, or Institutional Review Board (IRB) restrictions.
2. **Integrated structure.** Every item of ComNet data – both core and non-core – will have a unique and specifiable position within a single, comprehensive XML-based data grid. To achieve this tight integration, we will mesh existing formats into the single XML Scheme, thereby overcoming the Format Babel identified by Bird & Liberman (2001).
3. **Interoperability.** To promote interoperability, we will construct roundtrip conversions of ComNet data to other popular analytic programs for display and analysis.
4. **Powerful analysis.** Relying on ComNet XML, the project will build powerful tools for browsing, searching, visualization, statistical analysis, and report generation.
5. **Multilinguality.** ComNet will integrate data from hundreds of

languages, ranging from English and Spanish to minority and endangered languages such as Mapudungún, Ojibwe, and Iñupiaq.

4.2 Links across the sciences

ComNet will be based on the principles of open access and integrated structure. Open access to ComNet data will enable fundamental linkages between research in psychology, education, linguistics, biology, neuroscience, ethology, anthropology, sociology, political science, economics, demographics, computer science, natural language processing, human language technologies, library science, law, area studies, and comparative literature. The system will be capable of linking individual-level transcript data to related neurological, economic, medical, survey, social, and preference data, subject to privacy and IRB constraints. The construction of these additional linkages will be directed and controlled by the participants themselves.

4.3 Example projects

When ComNet resources are in place, it will be possible for users to address hundreds of types of research questions. Here are some illustrations of the types of questions that researchers have been asking, sampled from this much larger space:

1. Analyze the contents of newspapers in six different languages for their reaction to the terrorist attacks of September 11, 2001, using LIWC (Pennebaker et al. 2001).
2. Find out what children in varying social groups have to say about going to the doctor and how parents respond to these expressions (Steward & Steward 2006).
3. Track uses of 鄙人 (bi3ren2) or 在下 (zui4xia4) for "your humble servant" in Chinese texts from the Ming and Ching dynasties in order to understand the time course of changes in class structure (Brown & Gilman 1960).
4. Examine the extent to which high school classes in mathematics in eight different countries engage in the process of "accountable talk" (Michaels, O'Connor & Resnick 2008) and how this discussion plays out in terms of whole class involvement (Pea & Hoffert, 2007).
5. Tabulate and graph the frequency of verb over-regularizations such as *goed* and *runned* in child language samples between ages 1;6 and 3;10 (Marcus et al. 1992). Pass the matches to either Excel or R for further statistical analysis.
6. Compare videotaped story-telling and route descriptions in cultures that construct geocentric spatial terms, such as Tzeltal (P. Brown 2001) or Guugu Yimithirr (Haviland 1993) with those that use allocentric reference. Relate these to ways in which people describe geophysical features of their environment (Mark et al.1999).

7. Extend the analysis of Seyfarth & Cheney (1999) for vervet monkey vocalizations to vocalizations from groups of meerkats (Hollén & Manser 2007), using data currently in TalkBank.
8. Use scene segmentation and gaze alignment to measure how people maintain eye contact (Vertegaal et al. 2001) and postural direction during survey interviews (Groves 2004, Schober & Conrad 2006) when they are producing the disagreement signal *well* or its equivalent in French (*bien*), Hungarian (*hát*), German (*ja, na ja*) or Spanish (*pués*).
9. Link ComNet data to ALFRED (alfred.med.yale.edu) to locate possible genetic markers for indigenous groups whose languages lack recursive syntactic devices (Everett 2007, Hauser, Chomsky & Fitch 2002).
10. Extend the method of Mitchell et al. (2008) that predicts fMRI brain activity from the meaning of nouns to the prediction of brain activity from the reading of types of passages. Extend this method across languages.
11. Create a concordance of terms referring to sustainable energy generation across European and Japanese parliamentary debates and output a set of audio files including the sentences with these references. These materials will support a wider comparison of differences in national energy policies (Jacobsson & Bergek 2004).
12. Study the ways in which Supreme Court justices signal their intentions to vote on a given case to other justices (Johnson 2004) during the years 1957 and 2008. Link these signals to cognitive analyses (Ashley 1991).
13. Study the way bias and perspectives are expressed in social media web artifacts such as blogs or YouTube.

These are sample illustrations taken from a much larger space. Focusing just on the one field of child language, MacWhinney (2008b) lists 50 such research theme types, while the wider literature demonstrates the importance of at least 100 more. When we look at other fields, we find a similar diversity. In each field, however, the barriers to analysis are the same. Absence of published metadata makes it difficult to find, group, select, integrate and analyze appropriate data. Moreover, once located, there are often the barriers of Licensing and Format Babel. By lowering these barriers, ComNet will allow researchers to ask questions in ways that are currently impossible. By allowing researchers to address these new issues, ComNet will enable a transformation in the study of human communication.

4.4 Data curation

The construction of a system as extensive as ComNet requires careful attention to the curation of data through privacy protection, provenance and copyright protection, data validation, organization, metadata generation, and documentation.

Privacy protection: Over the course of 24 years, TalkBank, CHILDES, and LDC have aggregated the world's largest databases of spoken language materials. The fact that this work has never triggered a violation of privacy regulations is indicative of the attention we have paid to this issue. IRB committees across the U.S. are now using our standards in determining their approach to making spoken language materials available. These policies are available at http://talkbank.org/share and have gone through repeated cycles of discussion and refinement from communities as diverse as teachers, parents, aphasiologists, computer scientists, lawyers, and patient advocates. Some of the procedures involved include:

1. Online storage of IRB materials and releases from data contributors.
2. Removal and bleeping of last names and addresses.
3. Password and encryption protection of sensitive materials.
4. Repeated checking with contributors and sometimes participants to make sure that the current level of access is in accord with their wishes and those of the participants.
5. Contacting individual children when they become adults to verify that they wish to allow continued access to their data.
6. Targeted de-accessioning from the database of segments or whole transcripts that seem embarrassing or which could lead to identification of the participant.

Provenance and copyright protection: In addition to issues of confidentiality, ComNet will deal with issues of intellectual property, copyright, provenance, and authenticity. Both LDC and TalkBank publish all of the software we have produced under the GNU Library General Public License (GPL), or equivalent open source license, and we will continue to do so. We choose the GNU Library GPL because it maintains open access, while protecting use for research. By default, we leave the copyright on corpora with the original holders who have agreed to allow us to make their data accessible. In addition, LDC maintains agreements with each of its users that are compatible with agreements previously negotiated with each provider. For books and video, copyright is negotiated in accord with procedures established by the UDL (Universal Design for Learning). The UDL includes a collection of pre-1923 historical texts, which are out of copyright. The UDL has already negotiated copyright permission for every publication from the National Academy of Sciences.

Data validation: Both LDC and TalkBank conduct content validation at the time data are ingested. This process involves checking for transcription accuracy, file correspondence, and metadata entry. Format validation, which is run automatically by current TalkBank programs, applies at the level of the transcript, the corpus, and the metadata. Because the database is structured in XML, it is easy to run tools that validate the adherence of new contributions to the standard Schema. Data must pass through a roundtrip from CHAT to XML

and then back to CHAT (the ComNet transcript display format) with no validation errors. For corpora that are linked to media, each media time tag must correspond with a media file correctly stored on the streaming media servers. All ComNet data must fit exactly into this schema.

Organization: To guarantee proper functioning of the database, all data are encoded within a set of five isomorphic trees for:

1. Raw CHAT data,
2. XML CHAT files,
3. Media matching the transcripts,
4. Streaming media matching the transcripts, and
5. Commentary pegged to both transcripts and media.

For example, the transcript for the 10th session of the Yasmin corpus is located at data-orig /romance/es/Yasmin/10.cha. The XML is at data-xml/romance/es/Yasmin/10.xml. The media is at media/romance/es/Yasmin/10.mov. The four alternative compressions of the streaming media are on a streaming server at /CHILDES/romance/es/Yasmin/10/ and the documentation in HTML format is at /CHILDES/romance/es/Yasmin/10.

Metadata generation: The next step of curation involves the creation and maintenance of metadata for ISBN cataloging, DOI (Digital Object Identifier) generation, and OLAC (Open Language Archives Community) indexing. TalkBank files are currently registering automatically to OLAC (www.language-archives.org) and IMDI (www.mpi.nl/IMDI/) metadata systems. Metadata are important for smooth functioning of virtually all aspects of ComNet. This means that, as the project grows, we must continue to improve and extend the metadata set. As an example, let us consider the role that metadata will play in the curation and analysis of speech data. Currently, the /ae/ sound in Canadian English is shifting toward the /ao/ sound (Labov, Ash & Boberg 2006). In 40 years, this transition may be complete and it will then be difficult to correctly process earlier data without having metadata that indicates the time and place of recording and the dialect background of the speaker. This underlines the urgency of placing labels on data as soon as possible. For this sound change, it is also important to record metadata regarding speaker age, gender, geography, and education (Labov 2001). Collection of this sociolinguistic metadata supports a new trend in speech analysis that allows researchers to develop more coherent statistical models. The examination of speech in the context of the oral interviews that accompany major national surveys provides a unique opportunity to link rich metadata to detailed speech analysis.

Documentation: The final step of curation involves the writing of a readable PDF description of each corpus for inclusion in the corpus description manual. This manual is structured both as an independent, readable document and

as a set of individual, searchable descriptions with metadata fields. In addition, the manual for the XML coding system is linked to fuller documentation in a manual that describes the format in terms of the CHAT display.

4.5 Interoperability

ComNet will provide interoperability for data, metadata, and programs.

Data interoperability: ComNet is building on 25 years of work on data integration in the TalkBank framework. TalkBank data derive from 158 separate projects, each of which was eventually integrated into the single over-arching XML framework. During this process of integration, it was often necessary to extend the framework to represent new contrasts or distinctions marked in particular corpora. This process will continue within ComNet. Because the schema is a growing framework, TalkBank tools have been constructed to allow for repeated cyclic reformatting and validation of the whole corpus whenever a change is made to the schema. Given the goal of constructing a single, integrated database, a major goal for ComNet will be the integration of LDC materials to the ComNet XML standard. Once this is achieved, users of LDC data will have improved access to LDC materials and can use ComNet tools to process these data. This merger of the two systems will also encourage additional research groups, and even new user communities whose data formats differ, to become LDC members, thereby further strengthening ComNet sustainability.

Metadata interoperability: Computational linguists have devoted a great deal of attention to the development of systems for annotating the ontologies of human communication. Among the efforts in this direction, we should single out OLAC, IMDI, GOLD, and WordNet. ComNet will build on each of these systems. The OLAC metadata set (Simons & Bird 2008) is a subset of the larger IMDI set (www.mpi.nl/IMDI/tools). At a minimum, ComNet data will subscribe to OLAC. However, ComNet will also support IMDI validation, as a further option. The textual segments of ComNet will be curated using the TEI (www.tei-c.org) metadata set. The TEI framework extends beyond metadata to specific language tags. ComNet will also work to integrate these lower level TEI tags into the overall ComNet XML Schema. We will also work to bring the ComNet schema into accord with the developing GOLD (Farrar & Langendoen 2003) ontology (linguistics-ontology.org). This ontology uses the SUMO upper ontology (www.ontologyportal.org) which itself is in accord with the WordNet (wordnet.princeton.edu) framework. GOLD has been most fully elaborated in regards to the features of morphosyntax that encode gender, number, person, case, evidence, evaluation, modality, tense, mood, force, size, aspect, polarity, and voice. This ontology allows us to elaborate the current XML schema to correspond to the decomposition produced automatically by the MOR program for the %mor (morphosyntax) line.

Program interoperability: Our work on program interoperability is fairly advanced. We can now convert between every major display format in this field and the ComNet XML format. However, as ComNet moves into new areas, such as library collections and survey data, we will need to further extend interoperability by extending the ComNet XML schema and the various programming tools that rely upon it to integrate with TEI and other formats.

4.6 Search and visualization

ComNet will also seek to advance methods for search, discovery, and analysis. Users will be able to access ComNet materials through a single entry port interface, using a search engine that can take advantage of the structured, open XML databases we have created. The ComNet searcher will support multilingual, metadata-driven search, composition of regular expression searches, and direct playback of transcripts and media located through the search routines. The search program will implement all standard technologies for concordances, frequency counts, mean length of utterance, tagging, sequence analysis, etc. The searcher will also transmit data to standard statistical analysis in Excel, R, Matlab, and other programs for report generation and data visualization.

For analysis of sociolinguistic data, we will extend the DASLTran tool developed for TalkBank so that it supports search and visualization of hierarchical and tabular data, for example, session and subject-specific metadata. This new tool will also import and export formats used within the quantitative sociolinguistic research community such as NCSLAAP (ncslaap.lib.ncsu.edu), Excel, GoldVarb (http://individual.utoronto.ca/tagliamonte/goldvarb.htm), R (www.r-project.org), and Praat (www.fon.hum.uva.nl/praat).

ComNet will also develop systems for automatic language analysis, such as part of speech tagging, grammatical dependency analysis, and a variety of content analyses, such as Pennebaker's LIWC (Pennebaker, Francis & Booth 2001). In various targeted areas, ComNet will develop tools to build specific information-extraction engines. For example, in the NSF-sponsored "Mining the Bibliome", Mark Liberman and his colleagues created corpora of abstracts of biomedical texts annotated syntactically and for biomedical entities, e.g., *carcinogens*, and the various ways they are named, mentioned, abbreviated, and referenced. These data were then used to build and evaluate systems that automatically identified and classified such entities. The team that built these systems consisted of linguists, computer scientists, programmers, and content experts.

To supplement these tools for data analysis, we will build additional tools for data visualization. Let us mention two as illustrations: BungeeView and Commenter. BungeeView (bungeeview.com) allows the user to discover patterns in metadata across documents and videos. Commenter implements a system for creating collaborative commentary (MacWhinney et al. 2004) much like the DIVER system (http://diver.stanford.edu/contact.html). Commenter provides both a browser front-end and a server database back-end to link comments to the media and to each other. A prototype version of this system is linked to the

prototype browser at talkbank.org/browser. The system will allow users to categorize and access comments by type, data, user group, and claim status.

4.7 Multimodal and video analysis

ComNet will provide unique opportunities for the development of new methods for analysis of the multimodal aspects of human communication, including the use of sign language. Many segments of the database will provide high quality transcriptions linked to linguistic, video, and content annotations. Using this rich annotation foundation, researchers will be able to align gestures and postures with changes in scenes, gaps in conversation, and many other structural features of conversational interactions and narratives. The multimodal analysis community currently relies on four major annotation tools: Anvil, ELAN, EXMaRLDA, and The Observer. (Earlier, in Figures 3 and 4, we looked at an example of a TalkBank transcript from a problem-based learning class in medical school displayed in both CLAN and ELAN.)

In order to support efficient multimodal analysis, we must advance the state of the art in the area of video analysis and annotation. Here, we will rely heavily on work conducted in the Informedia Project (www.informedia.cs.cmu.edu). The following sections explain how this work within Informedia can be extended in the ComNet framework.

Video analysis data: The video analysis community has long attempted to bridge the gap from low-level feature extraction to semantic understanding and retrieval of the communicated content. To solve this fundamental problem, we will create a large shared video database as a focused target for further analysis and evaluation. This shared database will include media, transcripts, screen text data, web text metadata, corpus metadata, shot segmentation, image features (Gabor texture, Grid Color Moment, and Edge Direction Histogram), local feature descriptors, motion features (kinetic energy, optical flow, MPEG motion vectors), audio features (FFT, SFFT, and MFCC), and characterizations of the data with the LSCOM concept ontology. Over 60 TRECVID (Text Retrieval Conference Video, trecvid.nist.gov) participants have done this type of sharing of automatically extracted metadata for the non-public TRECVID collections. We will work with this community to create similar metadata for our open-access content.

Video annotation toolkit: To further bootstrap the process of annotation of the video test bank, we will provide a complete video annotation toolkit. This resource will allow researchers throughout the video community to annotate their own data, expand the concept ontology, and explore higher-level search services. We have already fielded several very effective systems, allowing annotators to efficiently label representative key frames in video, as well as longer video clips. We will make a robust version of this tool available to others early on in the project.

Video analysis toolkit: A number of researchers have expressed interest in applying our tools to their own data sets. Responding to this need, we will make key components of the Informedia library system available as open source, including shot detection, speech recognition, alignment, etc. This toolkit will also include modules for finding shots, labeling motions, and classifying content automatically. Using this suite, researchers can quickly customize tools, refine concept ontologies, and re-train classifiers for diverse applications. This allows further shared development of the software for analysis or services such as summarization, without having to expend many additional person-years in development.

Web-based annotation: We will also make use of the cyber-infrastructure opportunity provided by web games, such as the ESP game developed by CMU researcher Luis van Ahn (van Ahn, Kedia & Blum 2006), to allow collaborative annotation of video on the web. For this task we again expect the undergraduate students as well as the high school students participating in the summer programs at CMU to contribute ideas and implement code. This work builds on the notion of "human computation", whereby manual work is efficiently spread out over large numbers of people on the Internet, hereby providing innovative solutions for collaborative annotation of web video. Validation and verification of this annotation effort is done through duplication, which means annotations are only accepted after multiple people independently create the same label. This has been highly effective for image annotation and we will extend this annotation principle to our video collections.

Video annotation evaluation: Finally, to test the efficacy of new annotations against the video test bank, we will provide a benchmark set of tasks for video analysis evaluation.

4.8 Language communities

ComNet will also construct general methods for creating data and methods useful for specific language communities. We will work with four types of communities:

1. non-endangered U.S. minority languages, such as Spanish and Hawaiian.
2. endangered U.S. minority languages, such as Iñupiaq and Ojibwe.
3. non-endangered non-U.S. languages, such as Mapudungan, Welsh, Aymara, and Nahuatl.
4. endangered non-U.S. minority languages, such as Atayal, Cree.

The importance of work in language preservation for endangered languages is widely recognized. Without intervention, more than half of the world's 7,000 spoken languages are not expected to survive this century (Crystal 2000), and many others, even those with millions of speakers, are struggling to maintain a state of stable bilingualism within a surrounding dominant language community.

Many first language communities already have their own websites where group members can access culturally relevant materials. We will work with the developers of these sites to make them interoperable with ComNet tools and formats, while still making sure that communities maintain full control over their sites. Each site will have resources that will allow the target communities to construct social networking systems, cultural documentation, and links to community activities – all in the local language. The social networking software will be developed using open source versions of Facebook, role-playing games, and virtual worlds. These materials will rely on methods for linking transcripts to media, so that speakers can engage in direct dialogs over the web, while still producing written records of their conversations. Our language communities will also use facilities like the ComNet Commenter system to engage in blogging and commentary in the local language with all data being stored on the servers. Apart from these social uses of the web, we will also emphasize tools that assist in language learning and maintenance. These include online dictionaries and methods for grammatical analysis, as well as text-to-speech systems. We will also deploy local versions of the various language learning software methods developed by MacWhinney and colleagues at http://talkbank.org/pslc.

From the materials constructed by these communities, we will develop corpora for each language formatted in ComNet XML. These corpora will then be used for the development of orthographic and linguistic standardization, spelling checkers, online dictionaries, speech recognition and synthesis, information retrieval, telephone dialogue systems, and machine translation.

5. Conclusion

By including a focus on multimedia interactions linked to transcripts, corpus linguistics can vastly expand its horizons. This expansion will rely on several continuing developments. First, we need to develop easily-used methods for each of the ten analytic methods we have examined, including lexical analyses, QDA, automatic tagging, language profiles, group comparisons, change scores, error analysis, feedback studies, conversation analysis, and modeling. Second, we need to work together to construct a unified database for language studies and related sciences. In accord with the proposed ComNet structure, this database must be grounded on the principles of open access, data-sharing, interoperability, and integrated structure. It must provide powerful tools for searching, multilinguality, and multimedia analysis.

If we can build this infrastructure, we will be able to explore more deeply the key questions underlying the structure and functioning of language, as it emerges from the intermeshing of scores of processes operative on eight major timeframes.

References

Abiteboul, S., P. Buneman & D. Suciu (1999), *Data on the web: from relations to semistructured data and XML*: Morgan Kaufmann.

Ashley, K. (1991), *Modeling legal arguments*. Cambridge (MA): MIT Press.

Bates, E. & G. Carnevale (1993), 'New directions in research on language development', *Developmental review,* 13: 436-470.

Bates, E. & B. MacWhinney (1989), 'Functionalism and the Competition Model' in: B. MacWhinney & E. Bates (eds.) *The crosslinguistic study of sentence processing*. New York: Cambridge University Press. 3-73.

Bellugi, U., H. Poizner & E.S. Klima (1989), 'Language, modality and the brain', *Trends in neuroscience,* 12(10): 380-388.

Biber, D., S. Conrad & R. Reppen (1994), 'Corpus-based approaches to issues in applied linguistics', *Applied linguistics,* 15: 169-189.

Bird, S. & M. Liberman (2001), 'A formal framework for linguistic annotation', *Speech communication,* 33: 23-60.

Blanchard, D., J. Heinz & R. Golinkoff (2010), 'Modeling the contribution of phonotactic cues to the problem of word segmentation', *Journal of child language,* 37: 487-511.

Bohannon, N., B. MacWhinney & C. Snow (1990), 'No negative evidence revisited: beyond learnability or who has to prove what to whom', *Developmental psychology,* 26: 221-226.

Brown, P. (2001), 'Learning to talk about motion UP and DOWN in Tzeltal: is there a language-specific bias for verb learning?', in: M. Bowerman & S. Levinson (eds.) *language acquisition and conceptual development*. New York: Cambridge University Press. 512-543.

Brown, P., & S. Levinson (1987), *Politeness: some universals in language usage*. Cambridge: Cambridge University Press.

Brown, R., & A. Gilman (1960), 'The pronouns of power and solidarity', in: T.A. Sebeok (ed.), *Style in language*. Cambridge, MIT Press. 253-276.

Bybee, J. (2003), *Phonology and language use*. New York: Cambridge University Press.

Bybee, J. & P. Hopper (2001), *Frequency and the emergence of linguistic structure*. Amsterdam: Benjamins.

Corder, S.P. (1983), 'A role for the mother tongue', in: S. Gass & L. Selinker (eds.) *Language transfer in language learning*. Rowley (MA): Newbury House.

Crystal, D. (2000), *Language death*. Cambridge: Cambridge University Press.

Darwin, C. (1871), *The descent of man and selection in relation to sex*. London: John Murray.

Elman, J.L. (1993), 'Learning and development in neural networks: the importance of starting small', *Cognition,* 48: 71-99.

Enard, W., M. Przeworski, S. Fisher, C. Lai, V. Wiebe, T. Kitano, A. Monaco & S. Pääbo (2002), 'Molecular evolution of FOXP2, a gene involved in speech and language', *Nature,* 418: 869-872.

Everett, D. (2007), 'Challenging Chomskyan linguistics: the case of Pirahã', *Human development,* 50: 297-299.

Farrar, S., & T. Langendoen (2003) 'A linguistic ontology for the Semantic Web', *GLOT international, 7*, 97-100.

Freudenthal, D., J. Pine & F. Gobet (2010), 'Explaining quantitative variation in the rate of Optional Infinitive errors across languages: a comparison of MOSAIC and the Variational Learning Model', *Journal of child language,* 37: 643-669.

Friedl, J. (2002), *Mastering regular expressions*. Sebastopol (CA): O'Reilly.

Ganger, J. & M. Brent (2004), Reexamining the vocabulary spurt. *Developmental psychology,* 40: 621-632.

Gillis, S., W. Daelemans & G. Durieux (2000), '"Lazy learning": natural and machine learning of word stress', in: D. Broeder & J. Murre (eds.) *Models of language acquisition*. Oxford: Oxford University Press. 76-102.

Goldman-Eisler, F. (1968), *Psycholinguistics: experiments in spontaneous speech*. New York: Academic Press.

Goodenough, U. & T. Deacon (2006), 'The sacred emergence of nature', in: Clayton, P., & Z. Simpson (eds.) *The Oxford handbook of religion and science*. Oxford: Oxford University Press. 853-872.

Grosjean, F. & J. Miller (1994), 'Going in and out of languages', *Psychological science,* 5: 201-206.

Groves, R. (2004), *Survey errors and survey costs*. New York: Wiley.

Gupta, P. & D. Touretzky (1994), 'Connectionist models and linguistic theory: investigations of stress systems in language', *Cognitive science,* 18: 1-50.

Hall, W.S. & W.C. Tirre (1979), *The communicative environment of young children: social class, ethnic and situational differences*. Champaign, IL: University of Illinois.

Hauptmann, A., R. Yan, W-H. Lin, M. Christel & H. Wactlar (2007), 'Can high level concepts fill the semantic gap in video retrieval? a case study with broadcast news.' *IEEE transactions on multimedia,* 9: 33-39.

Hauser, M., N. Chomsky & T. Fitch (2002), 'The faculty of language: what is it, who has it, and how did it evolve?' *Science,* 298: 1569-1579.

Haviland, J.B. (1993), 'Anchoring, iconicity, and orientation in Guugu Yimithirr pointing gestures', *Journal of linguistic anthropology,* 3: 3-45.

Hay, J. & K. Drager (2007), 'Sociophonetics', *Annual review of anthropology,* 36: 89-103.

Hollén, L. & M. Manser (2007), 'Motivation before meaning: motivational information encoded in meerkat alarm calls develops earlier than referential information', *The American naturalist,* 169: 758-767.

Jacobsson, S. & A. Bergek (2004), 'Transforming the energy sector: the evolution of technological systems in renewable energy technology', *Industrial and corporate change,* 13: 815-849.

Jefferson, G. (1984), 'Transcript notation', in: J. Atkinson & J. Heritage (eds.) *Structures of social interaction: studies in conversation analysis*. Cambridge: Cambridge University Press. 134-162.

Johnson, T. (2004), *Oral arguments and decision making on the United States Supreme Court*. New York: SUNY Press.

Jones, M. & G. Conti-Ramsden (1997), 'A comparison of verb use in children with SLI and their younger siblings', *First language,* 7: 165-193.

Keenan, J., B. MacWhinney & D. Mayhew (1977), 'Pragmatics in memory: a study in natural conversation' *Journal of Verbal Learning and Verbal Behavior*, 16, 549-560.

Kendon, A. (1982), 'The study of gesture: some observations on its history', *Recherches sémiotiques/semiotic inquiry,* 2(1): 45-62.

Klatzky, R., B. MacWhinney & M. Behrmann (eds.), (2008), *Embodiment, ego-space, and action*. New York: Psychology Press.

Labov, W. (2001), *Principles of linguistic change. Volume 2. Social considerations*. London: Blackwells.

Labov, W., S. Ash & C. Boberg (2006), *The atlas of North American English: Phonetics, phonology and sound change*. Berlin: Mouton de Gruyter.

Larsen-Freeman, D. (2006), 'The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English', *Applied linguistics* (27): 590-619.

Lee, L. (1966), 'Developmental sentence types: a method for comparing normal and deviant syntactic development', *Journal of speech and hearing disorders,* 31: 331-330.

Leont'ev, A. (1947/1981), *Problems of the development of mind (English translation by M. Kopylova)*. Moscow: Progress Press.

Li, P., X. Zhao & B. MacWhinney (2007), 'Dynamic self-organization and early lexical development in children', *Cognitive science,* 31: 581-612.

Ling, C. & M. Marinov (1993), 'Answering the connectionist challenge', *Cognition,* 49: 267-290.

MacWhinney, B. (2004), 'A multiple process solution to the logical problem of language acquisition', *Journal of child language,* 31: 883-914.

MacWhinney, B. (2005), 'Can our experiments illuminate reality?', in: L. Gershkoff-Stowe & D. Rakison (eds.) *Building object categories in developmental time*. Mahwah (NJ): Lawrence Erlbaum Associates. 301-308.

MacWhinney, B. (2008a), 'Cognitive precursors to language', in: K. Oller & U. Griebel (eds.) *The evolution of communicative flexibility*. Cambridge (MA): MIT Press. 193-214.

MacWhinney, B. (2008b), 'Enriching CHILDES for morphosyntactic analysis',. in: H. Behrens (ed.) *Trends in corpus research: finding structure in data*. Amsterdam: Benjamins. 165-198.

MacWhinney, B. (2009), 'The emergence of linguistic complexity', in: T. Givon (ed.), *linguistic complexity*. New York: Benjamins. 405-432.

MacWhinney, B. (2010), 'Computational models of child language learning', *Journal of child language,* 37: 477-485.

MacWhinney, B., D. Fromm, A. Holland, M. Forbes, & H. Wright (2010), 'Automated analysis of the Cinderella story', *Aphasiology,* 24: 856-868.

MacWhinney, B. & J. Leinbach (1991), 'Implementations are not conceptualizations: revising the verb learning model', *Cognition,* 29: 121-157.

MacWhinney, B., C. Martell, T. Schmidt, J. Wagner, P. Wittenburg, H. Brugman & E. Hoffert (2004), 'Collaborative commentary: opening up spoken language databases', *LREC 2004*. Lisbon: LREC. 11-15.

Malvern, D.D., B.J. Richards, N. Chipere & P. Purán (2004), *Lexical diversity and language development*. New York: Palgrave Macmillan.

Marcus, G., M. Ullman, S. Pinker, M. Hollander, T. Rosen & F. Xu (1992), 'Overregularization in language acquisition', *Monographs of the society for research in child development,* 57(4): 1-182.

Mark, D., C. Freksa, S. Hirtle, R. Lloyd & B. Tversky (1999), Cognitive models of geographical space. *International journal of geographical information science,* 13: 747-774.

Massaro, D. (1987), *Speech perception by ear and eye*. Hillsdale, NJ: Lawrence Erlbaum.

Maurice, K. (1983), 'The fluency workshop', *TESOL Newsletter,* 17: 429-436.

Menzel, R. & M. Giurfa (2001), 'Cognitive architecture of a mini-brain: the honeybee', *Trends in cognitive sciences,* 5: 62-71.

Michaels, S., C. O'Connor & L. Resnick (2008), 'Deliberative discourse idealized and realized: accountable talk in the classroom and in civic life', *Studies in philosophy and education,* 27: 283-297.

Mitchell, T.M., S.V. Shinkareva, A. Carlson, K-M. Chang, V.L. Malave, R.A. Mason & M. Just (2008), 'Predicting human brain activity associated with the meanings of nouns', *Science,* 320: 1191-1195.

Monaghan, P. & M. Christiansen (2010), 'Words in puddles of sound: modelling psycholinguistic effects in speech segmentation', *Journal of child language,* 37: 545-564.

Nation, P. (1989), 'Improving speaking fluency', *System,* 17: 377-384.

Nicoladis, E. & F. Genesee (1998), 'Parental discourse and code-mixing in bilingual children', *International journal of bilingualism,* 2: 85-99.

Ninio, A. & P. Wheeler (1984), 'Functions of speech in mother-interaction', in: C. Garvey, R. Golinkoff & L. Feagans (eds.) *The development of communicative competence*. Norwood (NJ): Ablex. 196-207.

Ochs, E. (1979), 'Transcription as theory', in: E. Ochs & B. Schieffelin (eds.) *Developmental pragmatics*. New York: Academic. 43-72.

Osgood, C.E. (1971), 'Where do sentences come from?', in: D.D. Steinberg & L.A. Jakobovits (eds.) *Semantics: an interdisciplinary reader in philosophy, linguistics, and psychology*. Cambridge: Cambridge University Press. 497-521.

Palmer, M., D. Gildea, & P. Kingsbury (2005), 'The Proposition Bank: an annotated corpus of semantic roles', *Computational linguistics*, 31, 71-105.

Parisse, C. & M.T. Le Normand (2000), 'Automatic disambiguation of the morphosyntax in spoken language corpora', *Behavior research methods, instruments, and computers,* 32: 468-481.

Pea, R. & E. Hoffert (2007), Video workflow in the learning sciences: prospects of emerging technologies for augmenting work practices. In R. Goldman, R. Pea, B. Barron & S. Derry (eds.) Video research in the Learning Sciences. Mahwah (NJ): Lawrence Erlbaum Associates. (427-460).

Pennebaker, J.W., M.E. Francis & R.J. Booth (2001), *Linguistic inquiry and word count (LIWC): a computerized text analysis program*. Mahwah (NJ): Lawrence Erlbaum Associates.

Perfors, A., J.B. Tenenbaum & E. Wonnacott (2010), 'Variability, negative evidence, and the acquisition of verb argument constructions', *Journal of child language,* 37: 607-642.

Piaget, J. (1954), *The construction of reality in the child*. New York: Basic Books.

Pienemann, M., B. Di Biase, S. Kawaguchi & G. Håkansson (2005), 'Processing constraints on L1 transfer', in: J.F. Kroll & A.M.B. DeGroot (eds.) *Handbook of bilingualism: psycholinguistic approaches*. New York: Oxford University Press. 128-153.

Presson, N. & B. MacWhinney (2010), 'The Competition Model and language disorders', in: J. Guendozi (ed.), *Handbook of language disorders*. New York: Routledge. 31-48.

Prior, A., & B. MacWhinney (2010), Beyond inhibition: a bilingual advantage in task switching. *Bilingualism: Language and cognition*, 13, 253-262.

Quinlan, P.T. (2003), *Connectionist models of development: developmental processes in real and artificial neural networks*. Hove, UK: Psychology Press.

Regier, T. (2005), 'The emergence of words: attentional learning in form and meaning', *Cognitive science,* 29: 819-865.

Rose, Y., B. MacWhinney, R. Byrne, G. Hedlund, K. Maddocks, P. O'Brien & T. Wareham (2005), Introducing Phon: a software solution for the study of phonological acquisition. In D. Bamman, T. Magnitskaia & C. Zaller (eds.) 30th Annual Boston University Conference on Language Development (pp. 489-500). Somerville (MA): Cascadilla Press.

Sagae, K., E. Davis, A. Lavie, B. MacWhinney & S. Wintner (2010), 'Morphosyntactic annotation of CHILDES transcripts', *Journal of child language,* 37: 705-729.

Scarborough, H.S. (1990), Index of productive syntax. *Applied psycholinguistics,* 11: 1-22.

Schegloff, E. (2007), *Sequence organization in interaction: a primer in conversation analysis*. New York: Cambridge University Press.

Schober, M. & F. Conrad (2006), 'Does conversational interviewing reduce survey measurement error?', *Public opinion quarterly,* 61: 576-602.

Selinker, L., M. Swain & G. Dumas (1975), 'The interlanguage hypothesis extended to children', *Language learning,* 25: 139-152.

Seyfarth, R. & D. Cheney (1999), 'Production, usage, and response in nonhuman primate vocal development', in: M. Hauser & M. Konishi (eds), *Neural mechanisms of communication*. Cambridge (MA): MIT Press. 57-83.

Simons, G. & S. Bird (2008), 'Toward a global infrastructure for the sustainability of language resources', Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation, Cebu City, Philippines. 87-100

Smith, J., M. Durham & L. Fortune (2007), '"Mam, my trousers is fa'in doon!": Community, caregiver, and child in the acquisition of variation in a Scottish dialect', *Language variation and change,* 19: 63-99.

Smith, L.B. & E. Colunga (2003), 'Making an ontology: Cross-lingusitic evidence', in: D.H. Rakison & L. Oakes (eds) *Early category and concept development: making sense of the blooming, buzzing confusion*. London: London University Press. 275-302.

Sokolov, J.L. (1993), 'A local contingency analysis of the fine-tuning hypothesis', *Developmental psychology,* 29: 1008-1023.

Steward, M. & D. Steward (2006), 'Children's conceptions of medical procedures', *New directions for child and adolescent development,* 1981: 67-83.

Thelen, E. & L. Smith (1994), *A dynamic systems approach to the development of cognition and action*. Cambridge (MA): MIT Press.

von Ahn, L., M. Kedia & M. Blum (2006), Verbosity: A game for collecting common-sense facts. Paper presented at the CHI 2006, Montréal.

Vertegaal, R., R. Slagter, G. van der Veer & A. Nijholt (2001), 'Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes'. CHI 2001, 3: 301-308.

Waddington, C.H. (1957), *The strategy of the genes*. New York: MacMillan.

Waterfall, H., B. Sandbank, L. Onnis & S. Edelman (2010), 'An empirical generative framework for computational modeling of language acquisition', *Journal of child language,* 37: 671-703.

Wiebe, J., T. Wilson & C. Cardie (2005), Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39, 165-210.

Yip, V. & S. Matthews (2000), 'Syntactic transfer in a Cantonese-English bilingual child', *Bilingualism: Language and cognition,* 3: 193-208.