Linguistic Patterns in Spontaneous Speech, 1-25
2009-1-025-001-000263-1

Can there be Standards for Spontaneous Speech? Towards an Ontology for Speech Resource Exploitation

Dafydd Gibbon
Universität Bielefeld

1. Speech resources, standards and spontaneous speech

By *speech resources* in this contribution means relatively homogeneous audio and visual speech corpora, including recordings, transcriptions, annotations, metadata, and perhaps associated word lists and corpus-based language models. The concept of *spontaneous speech* is in the centre of discussion in the following sections. The strategy taken in this presentation is to embed the characteristics of spoken language into a broader functional and structural linguistic context.

1.1 The problem

At first glance, the terms "standards" and "spontaneous" would seem to contradict each other. But spontaneity is complex and ambiguous, and speech which is spontaneous in one sense (not being read, not being consciously planned or rehearsed) may be quite non-spontaneous in another (in using sociolinguistically restricted codes, clichés, conventionalized phrases and idioms). The high dimensionality of the family of genres which might be called spontaneous speech is gradually becoming clearer as databases of spontaneous speech are being collected, and applications of resources—particularly in speech synthesis—are being made, and emotional speech and multimodal speech are becoming commonplace objects of investigation. And on the other hand, the standards referred to are metatheoretical guidelines for language and speech information interchange, not prescriptive instructions for human behaviour, though of course they may be interpreted as prescriptive specifications for speech technology application development.

In the present contribution, the main focus is to specify linguistic and phonetic background infrastructure for designing and using resources for spontaneous speech. Speech resources are generally purpose-built, whether they are intended for speech technology applications at one end of the scale or for conversation analysis at the other, but it has long been recognized that such data can have applications beyond the original purpose, and therefore mechanisms for sharing the data are required.

One such mechanism is ontology-based search and analysis. Ontologies in this sense are heuristic classification systems, including taxonomies, meronomies and other network structures, and were developed for inference in expert systems in Artificial Intelligence work in the 1980s. Recently, ontologies for language and speech have been developed, such as GOLD (General Ontology for Linguistic Description—Farrar & Langendoen 2003), and have become rather popular. But for the speech domain nothing comparable exists.

Assuming that such an ontology is just as necessary for speech—at the current state of search technology—as in other domains, an initial ontology, and surely a highly controversial one, is presented for discussion, concentrating on prosody (particularly pitch systems and timing) and disfluencies.

But although ontologies may be heuristically motivated and usually task-driven, repeated re-invention of the wheel in the area of speech resources for different task areas can be—again, heuristically speaking—rather a waste of time, energy and funding. So the strategy taken in the present contribution is to take applications-driven constraints into consideration, but to take a step back and look at more generic issues involved in the creation of a re-usable ontology.

1.2 Standards and spontaneity—a contradiction?

On the one hand, resources need to be standardized for information exchange, computational processing, and linguistic and phonetic analysis, otherwise they are of no use. At first glance, the terms "standards" and "spontaneous" would seem to contradict each other. But spontaneity is complex and ambiguous, and speech which is spontaneous in one sense (not being read, not being consciously planned or rehearsed) may be quite non-spontaneous in another (in using sociolinguistically restricted codes, clichés, conventionalized phrases and idioms).

The high dimensionality of the family of genres which might be called spontaneous speech is gradually becoming clearer as databases of spontaneous speech are being collected, and applications of resources—particularly in speech synthesis—are being made, and emotional speech and multimodal speech are becoming commonplace objects of investigation. Standards, on the other hand, reduce basic parameters and values to a manageable set, and are not intended to be completely comprehensive, whether institutional standards such as ISO standards, or *de facto* industry standards such as email formats, popular operating systems, or computer types, or academic standards such as the International Phonetic Alphabet for transcription (see Gibbon et al. 1997, Gibbon et

al. 2000 for discussion), or the EUROTYP[1] or WALS[2] category sets for descriptive linguistics (Haspelmath et al. 2005).

The purpose of this paper is not really to provide a definitive overview of the field. This would be a useful task, and coherent consolidation of previous results is very necessary. The purpose is more forward-looking, projecting experience in a number of projects and personal research ventures into a research space for the future. Furthermore, in view of the overall context of this contribution, the aim is to point towards underlying linguistic (descriptive, formal, computational) requirements specifications for a re-usable ontology of spontaneous speech, rather than towards operational requirements for a specific ontology of spoken language systems which are concerned with particular aspects with spontaneous speech or particular speech technology applications in this area. For these two reasons, the references will also be somewhat selective.

1.3 Overview

Section 2 is concerned with the specification phase of developing an ontology for spontaneous speech. Section 3 is concerned with the concept of ontology, and current contributions to discussion in this field. Section 4, starting from the basic premise that phoneticians, linguists, speech technologists all deal with the domain of signs, looks at linguistic and semiotic requirements for such an ontology, with particular attention to the adequacy of basic models of sign structure. In section 4, a brief overview of selected current discussion on the development of ontologies for spoken language (and language in general) is given. Section 5 introduces a generic structure for a re-usable ontology, which takes the content, the structure and the rendering of signs into account. Section 6 examines contributions of speech specific categories.

2. Specifying resources for spontaneous speech

This section is concerned with taking a close look at specifying resources for spontaneous speech. Speech resources are generally purpose-built, whether they are intended for speech technology applications at one end of the scale or for conversation analysis at the other, but it has long been recognized that such data can have applications beyond the original purpose, and therefore mechanisms for sharing the data are required.

[1] http://wwwlot.let.uu.nl/Research/ltrc/eurotyp/index.htm
[2] http://wals.info/

2.1 On defining "spontaneity"

It is a long time since the Text Encoding Initiative produced initial recommendations for the transcription of speech corpora; in fact this was before the blossoming of the linguistic resources and language documentation paradigms, and before the development of computational methods of dialogue modelling (recently adaptations for XML have been made, but without substantive extension). So let us take a step back and ask a basic question: what do we mean by "spontaneous"?

The Merriam-Webster online English dictionary provides the following definitions:

Etymology: Late Latin spontaneus, from Latin *sponte* of one's free will, voluntarily

1: proceeding from natural feeling or native tendency without external constraint
2: arising from a momentary impulse
3: controlled and directed internally: SELF-ACTING <spontaneous movement characteristic of living things>
4: produced without being planted or without human labor: INDIGENOUS
5: developing or occurring without apparent external influence, force, cause, or treatment
6: not apparently contrived or manipulated: NATURAL

The term is evidently highly polysemous. Nevertheless, all these readings point in the same direction as the term *authentic* as it is currently used in foreign language teaching methodology. In this context, authentic texts are simply texts which are not produced for the purpose of language study.

In speech technology, linguistics, phonetics and psycholinguistics, a number of definitions have traditionally been offered, which were summarized in a *Linguist List* discussion over a decade ago (Fagyal 1995), which has lost none of its relevance:

'Spontaneous speech' is a
(1) type or 'mode' of speech production opposed to 'read-aloud' speech;
(2) real-time generated, unplanned and non-rehearsed type of encoding linguistic information;
(3) casual 'way of speaking' or 'style', characterizing informal speech situations;
(4) naturally occurring, non-experimental type of speech event of any kind.

These definitions are clearly narrower. The second and the fourth definitions correspond to the general dictionary definition of *spontaneous*, the third is the one which a

linguist concerned with language varieties would choose. The first definition is, however, probably the definition which is most common in the phonetic, psycholinguistic or speech technological laboratory. The first definition is a compromise, and very incomplete. There is, after all, such an activity as spontaneous read-aloud speech, for instance when one reads an extract from some text, whether a newspaper or a menu, to a companion.

2.2 Spontaneity or authenticity?

As already noted, the closest definition to what we need appears to come from foreign language teaching, which of course has centuries of experience with which the few decades during which speech technology has been around cannot compete. So let us take the notion of *authentic text* and its spoken language twin *authentic speech* to be what we are looking for:

> *Authentic speech is speech which is not produced for the purpose of the study of speech.*

This definition should serve us in good stead, as long as we bear in mind that for many purposes laboratory speech is *authentic laboratory speech*...

But now we have a problem: authentic speech is simply everything, and thus needs delimitation. One of the delimitation strategies which have been used during the past few years is to use *emotional speech*. But consider the range of uses of spoken language—from spontaneous discussions among academics to motherese between mothers and children, and from chance conversations between strangers to everyday talk between married couples. All of this can be classified as authentic speech, as spontaneous speech, but the main classification of these speech registers or styles is rarely "emotional vs. unemotional" speech. A further problem here is that so-called "emotional speech" is often, rather, emulated emotional speech based on stage conventions learned by professional actors. So we need to look for a parametric space within which speech instances are located.

3. Linguistic requirements for an ontology of signs

One strategy for restricting the parametric space for domain descriptions is ontology-based search and analysis; this section is concerned with characterizing the term "ontology", and in exploring linguistic models of language structure in search of an appropriate set of foundational categories.

3.1 On defining "ontology"

The term "ontology" is ambiguous. A general definition of *ontology* in its traditional sense can be found in Floridi (2003:155):

> Ontology as a branch of philosophy is the science of what is, of the kinds and structures of the objects, properties and relations in every area of reality.

And of course as empirical scientists, we find this notion somehow appealing, until we realize that *reality* is itself an elusive concept. A homogeneous definition is perhaps even less easy to find in contemporary computational usage, however. Floridi (2003: 158) provides a very general definition:

> In the field of information processing there arises what we might call the Tower of Babel problem. Different groups of data-gatherers have their own idiosyncratic terms and concepts in terms of which they represent the information they receive. When the attempt is made to put this information together, methods must be found to resolve terminological and conceptual incompatibilities. Initially, such incompatibilities were resolved on a case-by-case basis. Gradually, however, it was realized that the provision, once and for all, of a common backbone taxonomy of relevant entities of an application domain would provide significant advantages over the case-by-case resolution of incompatibilities. This common backbone taxonomy is referred to by information scientists as an 'ontology'.

The term *taxonomy* is not meant here in the restricted sense of lexical semantics: a hierarchy of classes or categories which is defined by relations of implication. It is meant more in the sense of *semantic network*, that is, a system of categories linked by several kinds of relation, including that of implication, but also including part-whole relations, temporal and spatial relations, and other kinds of relation.

Ontologies in this sense are classification systems, including taxonomies and other network structures, and were developed for inference in expert systems in Artificial Intelligence work in the 1980s.

The E-MELD project definition is useful (Anon 2005):

> An ontology here is essentially a machine-readable formal statement of a set of terms and a working model of the relationships holding among the concepts referred to by those terms in some particular domain of knowledge. Its purpose is not to define meaning, but to allow computers to navigate human knowledge

in a way that mimics intelligence.

But let us de-mythologize the term even more thoroughly:

> An ontology is a highly structured terminological dictionary designed to facilitate search for information in some technical domain.

It should be emphasized that in general ontologies are heuristically motivated: they do not impose constraints on new theoretical requirements which may arise, but need to be flexible enough to accommodate such requirements for practical search purposes at some level of conceptual granularity.

3.2 GOLD: General Ontology for Linguistic Description

The most well-known ontology for linguistics is currently the *General Ontology for Linguistic Description* (GOLD[3]), developed by Farrar & Langendoen (2003) as part of the metadata standardization effort of the E-MELD[4] (Electronic Metastructures for Endangered Languages Data) project. This ontology, and its motivation, is a good example of why ontologies are important for scientific activities: the more resources there are, the harder it is to search for and find relevant resources, and the more necessary it is to describe resources in a heuristically useful agreed vocabulary. But for the speech domain nothing comparable exists. The essential features of GOLD are conveyed in this definition from the GOLD website:

> GOLD is an ontology for descriptive linguistics. It gives a formalized account of the most basic categories and relations (the "atoms") used in the scientific description of human language. First and foremost, GOLD is intended to capture the knowledge of a well-trained linguist, and can thus be viewed as an attempt to codify the general knowledge of the field. GOLD is aimed at facilitating automated reasoning over linguistic data and at establishing the basic concepts through which intelligent search can be carried out.

In its current state, GOLD has a number of weaknesses:

1. The methodology is heavily slanted towards particular structuralist and generative descriptive methodological traditions, and is acknowledged to be in need of

[3] http://emeld.org/gold
[4] http://emeld.org/index.cfm

extension to include functionalist or diachronically oriented categories and relations.

2. In terms of linguistic unit size, GOLD is restricted to the traditional word constituents of phoneme and morpheme, and to sentence constituents. Text and dialogue modelling, which are crucial for an adequate ontology for spoken language, is not included.
3. The domains of inter-modality relations, in particular the role and structure of prosody and conversational gesture, are not included. Indeed, phonology and phonetics in general are also not well represented, though initial proposals have been made by Aristar (2005) for phonetics and Kamholz (2005) for phonology.
4. More fundamentally, GOLD pays little attention to the functionality of structures and their constituents. The functions of spoken language and its compositional or idiomatised parts, are at the core of an integrated theory of language.
5. Finally, as with other current ontologies, GOLD is biassed towards the analysis of text—whether text in the sense of written language, or text in the sense of traditional linear phonetic and phonological transcriptions, whereas spoken language requires compositional operations of overlap (also known as association, alignment, or parallelism) in addition to concatenation.

The last point is worth dwelling on a little. Formally, written texts are analyzable by means of concatenation grammars, at least until the domain of text and *Document Type Description* is reached. Concatenation is also adequate for the description of strings of phonemes, morphemes, words, sentences and so on, and can be used to define syntagmatic hierarchies over sequences of units (i.e. constituent, part-whole hierarchies, as opposed to classificatory taxonomic hierarchies). In the context of texts, concatenation can be interpreted as immediate proximity within a stylized spatial coordinate system. In the context of speech, concatenation can be interpreted as immediate proximity within a stylised temporal coordinate system with a relation of *temporal precedence* (corresponding to concatenation).

But also, in speech, a relation of *temporal overlap* is also needed, in order to express prosodic or "supra-segmental" and gestural facts of many kinds, from vowel harmony through intonation to the co-occurrence of emphatic accentuation and emphatic gestures. The range of overlap functions is broad, and may be based on "infra-segmental" structures (as in assimilation patterns) or on relatively independent "autosegmental" structures (as in tonal and intonational patterns or co-occurrent conversational gesture). In fact, a *spatial overlap* relation also holds for writing systems in the visual domain, in terms of the "prosody" of writing, such as highlighting and layout, though this is rarely dealt with in linguistic studies.

3.3 Linguistic essentials for a spontaneous speech ontology

Not all of the points listed as weaknesses of the current version of the GOLD model can be dealt with in the present contribution. The main points covered are:

1. Formal, structural features of spoken language, with particular reference to rank hierarchical structure and to prosody.
2. A ternary model of the relation between signs and the world, the *Content-Structure-Rendering* (CSR) model, which differs from traditional dualistic sign models in introducing an intermediate component of Structure, and in adopting a dual interpretation of Structure in terms of the world of Content and the world of Realization, following modern linguistic theories.

By Realization is meant the form of signs in different modalities, for example the acoustic realization of signs as speech sounds, and the visual realization of signs as gestures or as text.

In the following subsections, an outline of an initial taxonomy is established, and in the following sections functional aspects and the CSR model are outlined.

3.3.1 Linguistic categories with prosodic relevance

1. Basic category: speech event, a pair of a category and an interval
2. Structural levels of analysis, ranks (units of increasing size)
    1. Phoneme/toneme; syllable
    2. Morpheme/morphophoneme/morphotoneme
    3. Word (simplex, derived, compound)
    4. Phrase, sentence
    5. Text
    6. Dialogue
3. Semantic and pragmatic interpretation (for “concept annotation”)

3.3.2 Linguistic relations

1. Syntagmatic relations in speech:
    1. Sequential (concatenative and hierarchical) relations
    2. Parallel (autosegmental association) relations, including synchronization issues (“absolute slicing”, phonetic operations such as assimilation)
2. Paradigmatic (classificatory) relations of similarity and difference:
    1. Dependent on classification by sequential relations

    2. Dependent on classification by parallel relations
3. Interpretation relations:
    1. Manifestation relations (modality/media oriented):
        1. Acoustic
        2. Visual
    2. Content relations (semantics/meaning/function oriented):
        1. Contrast (phonology, Asian and African tonology)
        2. Morphemic (morpho-syntax, African tonology)
        3. Text, discourse

4. Semiotic bases for a re-usable spoken language ontology

Speech is more than arbitrary patterns of sound; to be speech, the patterns need to be accepted behaviour of a community, and they need to be interpretable in regular ways in terms of the needs of the community for information and action. This section is concerned with the embedding of basic structural concepts into functional frameworks, and outlines a number of traditional approaches for this purpose.

4.1 The Saussurean dualist conceptualist model

The first modern linguistic approach to modelling speech, both from a structural and a contextual perspective, was by the father of modern linguistics, Ferdinand de Saussure. A representation of his model, which is mentalist, in conceiving the sign in terms of mental thoughts and images, and also dualist, in structuring the sign into two parts, the *concept*, concept or thought, which refers to the meaning or *signifié* ("signified") of the sign and the *image acoustique*, acoustic image, which relates to the form or *signifiant* ("signifiant") of the sign. The model is shown in Fig. 1.

A further interesting feature of the model (which is often forgotten in the literature) is that de Saussure's mentalism is of a specific kind: the mind is understood as a "collective subconscious", coordinated by means of a circuit of sign exchange between members of the speech community, and thus introducing the notion of a *channel* or contact between interlocutors; for the significance of the *channel* for the functionality of prosody cf. Gibbon (1976).

The dualist approach provides the minimum number of components for a re-usable ontology which takes the notion of sign seriously.

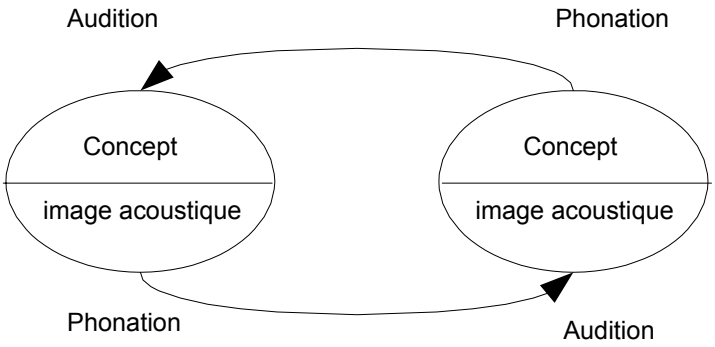

Figure 1: Dualist conceptualist model (de Saussure)

4.2 The Praguean functionalist constitutive factor model

A decade and a half later, in 1934, the psychologist Karl Bühler developed a four-component modification of the Saussurean circuit model in which the *sign* component was abstracted away from the individual interlocutors and presented as an entity which stood in a functional relationship with other components of the speech situation, the transmitter, yielding the expression function (Ausdrucksfunktion), the receiver, yielding the appeal function (Appellfunktion), and the context, yielding the representation function (Darstellungsfunktion) of language, as illustrated in Fig. 1.

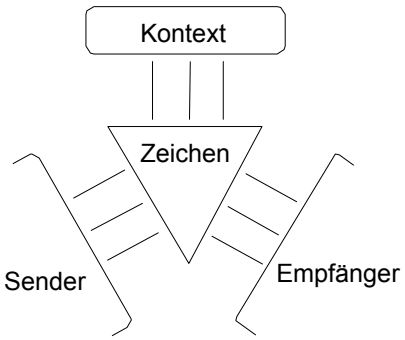

Figure 2: Functionalist instrumental model: Zeichen = sign,
Sender = transmitter, Empfänger = receiver, Kontext = context (Bühler).

Almost 30 years later, in 1960, an Praguean extension of Bühler's model to six components was presented by Roman Jakobson, in which sign was renamed "message", and additionally the channel was made explicit as the "contact" (not only the physical

channel, but also the communicative bond between the interlocutors), and the code. Like Bühler, Jakobson related the functions of language to the components of the communication situation, which he termed "constitutive factors": expressive (Sender), conative (Receiver), representational (Context), corresponding to Bühler's components, and metalingual (Code), poetic (Message) and phatic (Contact). The model is shown in Fig. 3. As these models were developed, more situational factors were introduced which are relevant for the description of spoken language, and in particular prosody: the metalingual function subsumes the configurative and delimitative functions of prosody in relation to locutionary structures (cf. accent positioning, contours, and boundary tones), and the phatic function has already been referred to in connection with calling intonations (Gibbon 1976).

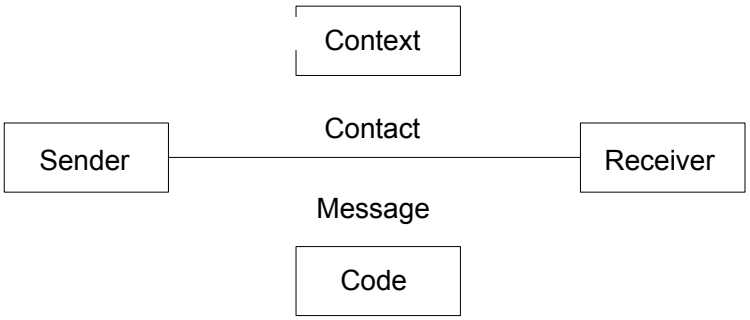

Figure 3: Functionalist constitutive factors model (Jakobson)

4.3 The Carnapian inclusion hierarchy model

A model with a different perspective was introduced by Rudolf Carnap (e.g. Carnap 1958:79), which has remained a kind of standard model in logic and linguistic semantics and pragmatics. The study of the structure of the sign, syntax, is embedded in the study of the meaning of the sign, semantics, which is in turn embedded in the study of the use of the sign by interlocutors, pragmatics. This model is not unrelated to the constitutive factor models, but basically prioritizes the components; it is no accident that Carnap and Bühler were contemporaries in Vienna. The syntax of the sign, i.e. the grammar of the forms and structure of the sign is the formal basis for a full description of the 'semiotic' of the sign, in Carnap's terminology. The context, with the representational function, semantics, is more encompassing, while pragmatics, encompassing the speaker and the hearer, provides the comprehensive environment for the other components.

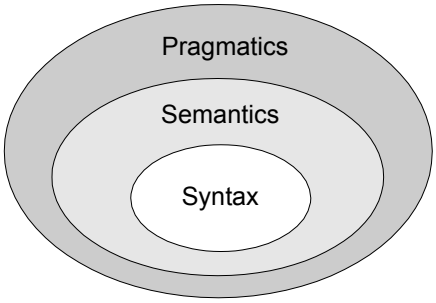

Figure 4: Pragmatic-Semantic-Syntax Inclusion model

It is noteworthy that while the other linguistic models gave much attention to syntax, with the exception of Jakobson, the introduction of syntax into the model, the internal structure of the sign, still does not do justice to essential features of spoken language: pronunciation, i.e. sounds, syllables, and the prosodic hierarchy.

4.4 The system cascade model

The Carnapian inclusion model has been re-interpreted in countless approaches to language modelling in the human language technologies as a cascade: for generation, pragmatics comes first, semantics follows, and syntax is the final stage. Again, there is no place for essential components required in a re-usable ontology for spoken language, from concepts of ritual, routine and idiomaticity, through the lexicon to the written-spoken distinction itself and, with it, phonetic interpretation and the prosodic hierarchy. The cascade model in its simplest form is shown in Fig. 5.

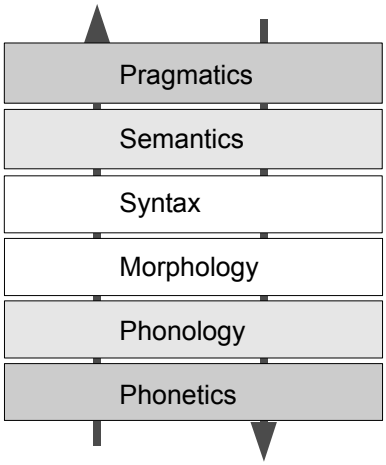

Figure 5: Cascade model of spoken language system architecture

The idea behind this model is that speech production starts with pragmatics, runs through semantics, syntax, morphology and phonology until it reaches the phonetic output. Conversely, speech perception and understanding starts with phonetics and proceeds in the other direction through phonology, morphology, syntax and semantics until the full pragmatic interpretation in context is reached.

This model is very useful for many purposes, and has been used on many occasions, in this or in more elaborated form. But from the point of view of a re-usable ontology it is flawed in more than one respect:

1. The relations between the components are very different: while sentences may be said to be composed of words (syntax), and words of morphemes (morphology), and, though more indirectly, that morphemes are composed of phonemes (phonology), it is not at all obvious that phonemes are composed of phonetic units, such as features or articulation phases in the same way. Neither is it obvious that pragmatic units are composed of semantic units and that semantic units are composed of syntactic units in the same way. Different relations are involved.
2. Syntax and morphology pertain to the structure of the sign, and are very different kinds of entity from semantics and pragmatics, which relate the sign to its real-world environment, and from phonology and phonetics, which give the sign its tangible real-world form.
3. The model is sentence and word oriented, and implies compositionality. There is no room for higher levels of structure such as text and dialogue, which are, in some indeterminate way, consigned to a catch-all pragmatics component.
4. There is no room for a distinction between compositionality of signs and idiomatization or routinisation of signs, which (among other things) are important for the assignment of intonation contour types.
5. There is no room for the incorporation of a prosodic hierarchy, in the sense of a mapping of locutionary units of different sizes into configurational trajectories of phonetic features of different sizes, for instance major and minor intonation phrases, foot units, syllables.

5. The CSR model: content, structure and rendering

The critique of the cascade model in the previous section provides the foundation for a more viable and differentiated model, which is likely to serve better in a re-usable ontology of spoken language. The present section discusses the ternary Content-Structure-Rendering model before the background of *co-interpretation* of signs in terms of the

world of content and the world of appearances of signs. An initial discussion of the lexicon in terms of this ternary model can be found in Gibbon (2002).

5.1 Semiotic co-interpretation: the semantics of domains and modalities

A set of core formal metatheoretical concepts will be introduced here. The three part-whole relations of *precedence*, *overlap*, and *hierarchy* constitute what are known as *syntagmatic relations* (structure-building) in linguistics, in contrast to *paradigmatic* or *classificatory* relations (of similarities and differences between categories). Both these types of relation have been formalized in so-called *feature grammars* as attribute value pairs. Attributes are in general taken to represent functional parts of a larger structure which is represented by an attribute-value matrix, while sets of values of attributes represent the partitions and equivalence classes which characterize paradigmatic relations.

It is this complex of relations which determines the *levels of analysis* in linguistics, for example the realization of phonemes as allophones, of morphemes by phonemes, of (some) morphosyntactic categories by tone in tone languages, and of utterance and discourse categories by intonation. Formal accounts of this complex of relation are given in Generative Phonology, in phonological Optimality Theory, Declarative Phonology, Computational Phonology, and Two-Level Morphology, for example. A theoretically founded and operational computational model of part of this complex of relations is to be found in the *Time Map Phonology* of Carson-Berndsen (1998).

The structure of an elementary sign with the co-interpretation architecture is shown in Fig. 6.

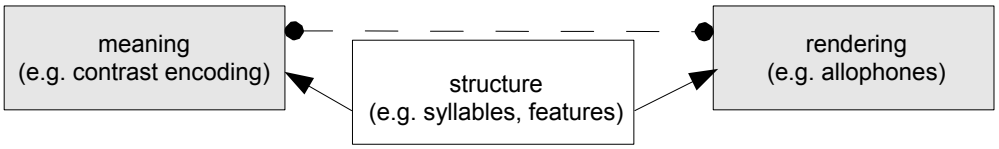

Figure 6: Sign structure with domain and modality co-interpretation

5.2 Co-interpretation in the hierarchical CSR architecture

The architecture of the CSR model is illustrated in Fig. 7. The core of the model is a hierarchical model of *ranks* of signs which not only differ in size but also in functionality. At each level, local kinds of hierarchy are also present: syllable structure, morphological structure, phrase structure, text structure and dialogue structure.

The sign units are represented in the middle column of the figure, and constitute a

rank of inclusive units with inclusive functions. At each rank, there are two interpretations, one, domain interpretation, in terms of content or function, and the other, modality interpretation, in terms of rendering or appearance. The sign is therefore an abstract (or mental) unit which relates in these two ways to reality: the content domain and the modality domain. The linguist can discourse about any part of this model, of course: in this metalinguistic discourse, the entire model—content, structure and rendering—then becomes part of the metalinguistic content domain, so in a sense the modality domain is, at a very general level, always potentially a part of the content domain in metalinguistic discourse.

With regard to the content domain, there are many appropriate theories of formal and functional semantics, pragmatics, conversation description and modelling which are easily accessible and which are not of concern here. In the centre of attention is the specific formal structure of spoken language.

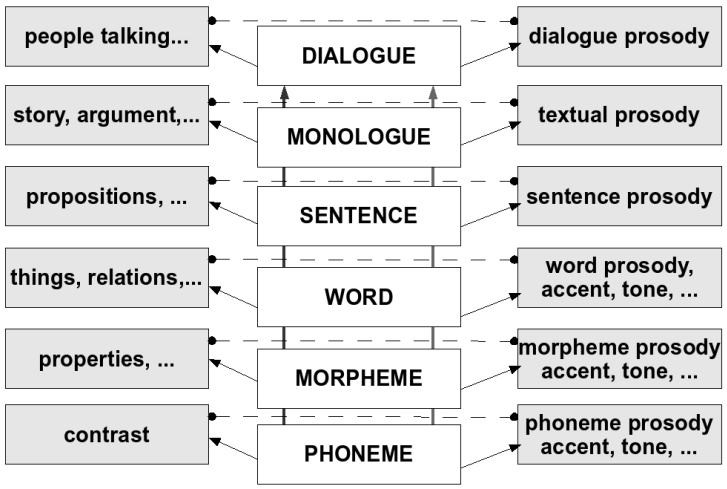

Figure 7: Content-Structure-Rendering (CSR) architecture

5.3 Structure: metasyntactic meronomies and taxonomies

These very basic relations, which are to be found in much introductory linguistic literature, but are rarely made explicit either in theoretical or applications work on linguistic matters, are primary requirements for a linguistic ontology for speech. Summarizing, the basic relations which determine linguistic levels of analysis required for describing speech resources are:

1. Syntagmatic relations (sequential and parallel) determining part-whole relations in complex constructions.
2. Paradigmatic relations (categories, classes) determining similarities and differences.
3. Interpretation (manifestation) relations (the modality and semantic-pragmatic interpretations) determining time types.

Failure to recognize these elementary distinctions has led to much confusion in phonological theory over the years, but in a workable linguistic ontology the distinctions are essential at all levels of description.

5.3.1 Syntagmatic (meronomic) relations

1. In syntax, sequential syntagmatic relations are expressed by labels such as *SUBJECT*, which represent sequential parts within a larger attribute-value matrix; overlap relations are generally ignored, but would be needed in accounts of intonation, accentual focussing, and morphosyntactic tone.
2. In phonology and prosody, sequential syntagmatic relations are expressed by labels such as *ONSET* (of syllables), *NUCLEUS* (ambiguous—of syllable or of intonation group), *ACCENT* (associated with syllable nuclei), which represent either sequential parts (syllable *ONSET* or *NUCLEUS*, intonation *NUCLEUS* and *ACCENT* with respect to other parts in intonation structure) or overlapping parts such as *ACCENT* with regard to syllable, word or sentence association.

5.3.2 Paradigmatic (taxonomic) relations

1. In syntax, paradigmatic relations hold over sets of contrasting items which enter the same syntagmatic relation, such as *pronoun*, *noun phrase*.
2. In phonology and prosody, paradigmatic relations hold over sets of contrasting items which enter the same syntagmatic relation, such as contrasting consonants and consonant clusters in *ONSET* position, or contrasting L* or H* accents in *ACCENT* position, or rising LH and falling HL pitch contours in *NUCLEUS* position.

5.4 Rendering interpretation: modality semantics

But this is not all. There is one other important complex of relations, namely the relation of linguistic structures to produceable and perceivable motor-sensory modalities,

which is variously referred to in different linguistic frameworks as *expression*, *realization*, *manifestation*, *interpretation* (e.g. *phonetic interpretation*). Generally this relation pertains to the acoustic modality, but if conversational gestures are included, as is becoming more and more common in many branches of linguistics and the human language technologies (Trippel et al. 2003), then the visual modality also has to be included here. A serious theory of writing (which does not yet exist) will also need to include the visual modality into its interpretation model.

Formally speaking, the *modality interpretation* relation is based on a *modality model*, which consists of the following two components (Carson-Berndsen 1998, Gibbon 1992, 2006):

1. a *domain* of phonetic categories and relations, including the *time type*—categorial, relational or absolute time types—and the *time map* between types as well as *precedence*/*overlap relations* which constitute type structure, and
2. a *function* which maps linguistic, phonological and prosodic units into the domain.

Nor is this all: different syntagmatic relations and the paradigmatic choices associated with them have to be interpretable in terms of their meaning or function in discourse, as well as in terms of the modalities of expression. In generative linguistics, as well as in formal logic, this dimension is known as *semantic interpretation* and is associated with a *semantic model*. For simplicity, but also because the traditional terms overlap considerably, I use "semantic" here to cover both conventional *semantics* and conventional *pragmatics*: consider the meanings of deictic categories, of speech act verbs, of the scope of conjunction or negation as marked by intonation and accentuation, in which semantics (briefly, concerned with *truth*) and pragmatics (briefly, concerned with *use*), as conventionally defined, overlap.

Formally speaking, the *semantic-pragmatic interpretation* relation is based on a *semantic-pragmatic model*, which consists of the following two components:

1. a *domain* of semantic and pragmatic categories and relations,
2. a *function* which maps linguistic units into the domain.

Formally, therefore, modality interpretation and semantic-pragmatic interpretation are very similar; the domains differ (though the modality domain is clearly one which we can also speak about, and therefore, strictly speaking, is also a subdomain of the semantic domain). Indeed, the fundamental sound-meaning relation is neatly explicated as *the pair of modality and semantic interpretations*.

6. The Time Type model of speech modality rendering

6.1 Grounded formal categories for transcription and annotation

The secondary formal categories are primarily concerned with the instantiation of the primary categories in terms of types of representation and interfaces (mapping functions) between them. Much could be written about this area, and the transcription and signal annotation systems which implement it, but attention will be restricted here to those relations which are essentially concerned with speech resource creation. More work is needed on these relations, so initially some of the basic concepts will simply be listed, before being explicated in more detail below:

- Transcription is the assignment of a segment of a speech event to a symbol (orthography, IPA, iconic).
- Annotation is a pair of a transcription label and a time-stamp (in the simplest, single-track case; cf. also multi-track annotation, hierarchical (tree-bank) annotation, and multi-stream annotation as with audio-visual recordings using appropriate alignment software).

Underlying these ideas is the insight that communication takes place in simultaneous channels and that these channels may be in different modalities:

1. *Vocal-acoustic*: speech
2. *Vocal-visual*: lip-reading
3. *Gesture-visual*: gesture
4. *Gesture-acoustic*: clapping, snapping, stamping

Functionally determined submodalities of the same modality also need to be defined, for instance in the vocal-acoustic modality the following:

1. *Locution*
2. *Prosody*
3. *Paralinguistics*

6.2 Time types as determinants of levels of analysis

The following time-oriented ontology outline for formal speech-specific categories and relations is modified slightly from Gibbon (2006), and is based on Gibbon (1992). First, three Time Types are needed as a basis for prosodic event alignment in the present analysis:

Absolute Time relates to signal-oriented phonetics, that is, to time points and intervals determined by calibrated physical measurement. For example, standard digital signal sampling techniques generate Absolute Time structures. In the Absolute Time domain, the quantitatively measured lengths of phones, syllables, etc., are important. Impressionistic phonetic judgements on length and tempo, as practised in phonostylistics and discourse analysis, may be seen as coarse-grained and uncalibrated quantitative measures.

Relative Time relates to 'interpretative phonetics', i.e. phonology and prosody, and defines intervals and other relations between points in time with no explicit assignment to Absolute Time. Relative Time characterizes the prosodic phonologies; the key relations are sequence, overlap and hierarchy, which are interpretable in terms of the Absolute Time domain.

Categorial Time relates to underlying lexical and grammatical levels, in particular to categories linked by algebraic operations such as concatenation. In the Categorial Time domain, there is only a notion of temporally uninterpreted structure; to include a notion of time, phonetic interpretations into the Relative Time and Absolute Time domains are required.

The three-level distinction between Time Types is supported by work in formal linguistic theory, in particular in Event Phonology, in Time Type Theory and in the Time Map Phonology approach to alignment theory (for further references, cf. Gibbon 2006).

6.3 Speech mining procedures

Key data-mining procedures in the exploitation of temporal annotations for speech resource creation can now be formulated in terms of the Time Types:

1. Analog-digital transformation in the signal sampling process, between two subdomains of Absolute Time.
2. Annotation as a mapping of the quasi-continuous digital domain of speech signals into the discrete Absolute Time domain of annotation intervals.
3. Induction of temporal structures from the discrete Absolute Time subdomain of annotation intervals to linear and hierarchical Relative Time structures.
4. Mapping of Relative Time structures to Categorial Time grammatical and discourse patterns.

6.4 Event alignment: streams, tracks, tiers

Each of the three Time Types is associated with its own specific range of sequential

and partially aligned parallel structures at different theoretical and heuristic levels of description. The relevant levels for the present study are distinguished as follows:

1. a set of parallel signal *streams* (time functions describing continuous or discrete sampled speech signals),
2. partially aligned with a set of parallel annotation *tracks* (time functions describing discrete, categorial sequences of events, as in a speech editor, for example, with sampled speech signal and parallel annotation tracks),
3. which are often derived from specific phonological *tiers* (linguistic constructs defining partially aligned trajectories through a feature space in Relative Time, as in autosegmental and other prosodic phonologies).

The "stream-track-tier" terminology is intended to keep apart clean ontological levels which are often indiscriminately labeled with terms like 'tier', 'track', 'level', 'layer', 'stratum', 'stream'.

The following more detailed terminological overview is based largely on the related models of Event Phonology, Time-Map Phonology, and Annotation Graph theory (cf. Gibbon 2006 for further details, including references).

6.5 Speech events and their representation

Speech event: A pair of an Absolute Time or Relative Time interval and a trajectory or pattern of values in some phonetic dimension, parameter or feature.

Examples:

- an interval of 120 ms and a phone segment, as a static or a dynamic time function in Absolute Time on an annotation track;
- an interval of 10 ms and a pitch value in an Absolute Time F0 stream;
- an interval of 0.0208333 ms (corresponding to 48 kHz sampling rate) paired with an amplitude value in an Absolute Time signal stream;
- a pair of a phonological segment and its phonemic or feature-based properties in Relative Time.
- a signal annotation $<<x_{min}, x_{max}>$, transcription>, where x_{min} and x_{max} range over points, *transcription* ranges over textual symbols, $<x_{min}, x_{max}>$ ranges over intervals ($x_{max} - x_{min}$ ranges over durations); cf. the following (Brazilian Portuguese) syllable annotation extracted from a Praat annotation file:

  xmin = 0.48473069812858305
  xmax = 0.6301876830002222
  text = "koN"

Transcription: The name of the pattern of an event.

Annotation: The name of an Absolute Time event, consisting of a set of pairs of transcriptions and either interval time-stamp pairs or point time-stamps.

Point: The undefined primitive for defining intervals as a pair of points (whether abstract points as in Relative Time, or clock time points as in Absolute Time), ignoring for present purposes the traditional discussion on whether points or intervals are primitives.

Time-stamp: The name of a point or a pair of points (an interval) in Absolute Time, i.e. a calibrated quantitative designation of a relative to some pre-defined initial point (the term 'tick' is used in digital music and virtual machine technology).

Examples:

- Mon Mar 28 13:32:30 BST 2005
- 321.5 ms

Absolute interval: The difference between two time points.

Examples:

- the subtraction of two time-stamps
- the time elapsed between two metronome beats.

Relative interval: A segment at an abstract phonological level, related to other intervals by relations of precedence and overlap. A relative interval has no absolute duration unless explicitly mapped into an absolute interval.

Example:

- The epenthetic [t] in English [prints] "prince" arises when the end of the nasal event interval of [n] precedes the end of the occlusive event interval of [n].

Time-Map: A function within one Time Type or between Time Types, mapping one temporal representation into another.

Examples:

- speech signal digitization (analogue signal sampling),
- annotation (aligns digital speech signal with event label sequence),
- phonetic interpretation (mapping of lexico-syntactic representation of speech forms into a phonetic representation.

This list is not complete, but is intended to serve as an initial orientation point for formal elements of a speech ontology.

6.6 Substantive categories

To discuss a detailed ontology for the substantive categories of speech would go

much too far in the present context. For segmental phonology, fundamental issues are outlined by Aristar (2005) and Kamholz (2005) in the context of E-MELD project discussions of the GOLD ontology.

Aristar details the following questions with respect to the GOLD ontology and his own proposal for segmental phonetic categories (minimally edited here):

- What is missing in the ontology?
- What is present that should not be there?
- How do we handle binary features such as coronal and dorsal, tense and high for consonants? Are these phonetic, or something else? Do we need a separate feature node for these?
- Have we made a proper distinction between phonetics and phonology? E.g. does the notion syllabic/non-syllabic properly belong here? What a syllable can be is reasonably decided by the phonology of a language, except in cases like semivowels (e.g. [w], [j]) which automatically become a vowel if they are syllabic, no matter what the language. So syllabicity is partly phonological, and partly phonetic.
- The ontology defines only membership of classes. For example, the IPA symbol [u] is a member of the classes Vowel, Back, Round, High, Vowel_Symbol. But there are facts about language which cannot be described in terms of class-membership. For example, can a voiceless sound ever be laryngealized? Can a voiced segment be ejective? These are essentially constraints between sounds. What constraints does a phonetic ontology need?
- There are sounds written as clusters in standard transcriptions, e.g. nasal clicks, multiple articulations like the labial-velar [gb]. Phonologically and phonetically, these seem to function as single sounds. But transcriptionally they function as clusters. What is the best way to handle these?

At a metatheoretical level, these issues are addressed in the present contribution; however, detailed solutions are still to be worked out.

One subset of substantive categories comprises the alphabets used for labeling segmental and prosodic categories:

1. *Segmental*: Kamholz (2005) essentially follows the International Phonetic Alphabet (but also points out the need for a notion of gradual or scalar feature values). The IPA is certainly the major candidate for phoneme-sized segmental labeling. It is essential in discussing the IPA for ontological purposes to make the following distinctions:

1. The IPA as a set of phonetic categories, defined by a set of feature matrices.
2. The IPA as a set of glyphs (font shapes) representing these categories, for which there are many (mutually largely incompatible) mappings to numerical codes and implementations, such as:
    1. Unicode (output oriented for publishing; problems with manual character input),
    2. Truetype fonts (widespread, highly inconsistent amongst each other),
    3. Metafont tools for LaTeX (easily modifiable),
    4. SAMPA representation in ASCII (the most useful for research-oriented computation).

2. *Prosodic*: no proposals have been made so far for a prosodic ontology, and there are several proposals, each emphasizing different category and relation sets, but evaluations of these with regard to the requirements of speech resources are still not available. The main candidates are:
    1. The IPA prosodic categories (with special glyphs).
    2. Hirst's IntSint relational categories of Hirst (Hirst & Di Cristo 1998), with special glyphs and ASCII representations.
    3. Gibbon's SAMPROSA compendium of symbols (cf. Gibbon et al. 1997, with ASCII representations).
    4. The ToBI symbol set (cf. Silverman et al. 1992, with ASCII representations).
    5. Other representations:
        1. The "tadpole" iconic representation of tonetic language teaching materials.
        2. Numerical representations of stress and pitch heights.

7. Conclusion

This contribution has outlined some of the essential features of an ontology for speech resource administration and search, including formal ontological categories, in some detail, and has pointed out strategies for developing substantive ontological categories.

Many open questions remain, including the following:

1. Content:
    1. Which prosodic category systems—one, more than one, all?—are appropriate for including in a *General Heuristic Ontology for Speech Technology* (let's call it *GHOST*).
    2. Important areas of prosody are not included, such as speech timing, including rhythm (cf. Gibbon 2006).
    3. The area of conversational gesture (Trippel et al. 2003) was mentioned, but not discussed in detail.
    4. A whole area has not been covered by the present discussion, namely performance problems of *disfluency* (cf. Tseng 1999) and *discourse particles* (cf. Fischer 2000).
2. Implementation:
    1. What is the most useful mapping of the GHOST categories and relations discussed in the present contribution into markup languages such as XML?
    2. How can a comprehensive GHOST system be incorporated into current annotation software, and into a semantic web oriented search system such as OLAC?

These questions will, for the moment, be left open for further discussion.

Dafydd Gibbon
Fakultät für Linguistik und Literaturwissenschaft
Universität Bielefeld
Postfach 100 131, D-33501 Bielefeld
Germany
gibbon@uni-bielefeld.de