

Perspectives in corpus annotation

Andreas Mengel

Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart

Abstract

In the areas of speech processing, annotated data differ considerably. The reasons are: Different levels of description (e.g., phonetics, syntax, semantics) which can be subdivided (e.g., perception research, psycholinguistics, articulatory phonetics in phonetics), different theoretical approaches (e.g., ToBI vs. INSINT in intonation), and different representation formats for the data. The least problem here - though fundamentally important - is that of different annotation representations in files. Even if it is possible to align and relate annotation for different levels of description, annotation in corpora is always static and only an instantiation of the current theoretical knowledge within a specific research area and certainly subject to further revision. Static aspects of corpus annotation must be combined with flexible annotation frameworks. A flexible annotation framework allows to deduce new annotation from existing tags. Deduction in this sense means that it is possible to formulate rules or hypotheses, test result and either adapt the rules or the annotation data. This paper argues in favour of the integration, reuse, and flexibility of speech annotation, describes necessary properties of software supporting these and presents an annotation environment in which first steps in this direction have already been made.

Preprint submitted to Elsevier Preprint

27 August 1998

1. Status quo

Speech corpora are needed for the development of theories: This general description can be applied to both speech technology and basic research in phonetics. The annotation of speech is a prerequisite for the training of statistical algorithms and large-scale testing of new hypotheses or rule sets in linguistics. Hence, the accessibility of annotated data is crucial for progress in speech science.

Despite the existence of large amounts of speech data in general, their accessibility is far from being satisfactory. Factors which impede broad access to existing data are described in the following sections.

Representations

The most obvious obstacle for general access to corpora is a lack of standardisation of formats. *Format* in this sense denotes the grammar needed to parse a given corpus file and interpret the syntactic relations of its entities. A description and review of different annotation formats and their application can be found in [1].

Software

As an obvious consequence of the variety of different formats used in the field, different software packages are needed to read and write these formats. Additionally, all software and special routines for the manipulation and analysis are only available for corpora in the special format they have been implemented for.

Tags

The relation between a sign and its meaning is arbitrary and the same linguistic label can be represented by different tags in different corpora: The start time label for an utterance can be encoded by *start*, *s*, or *beg*, the time value can be provided in minutes, seconds, milliseconds, or even in samples.

Levels

Traditionally, human speech is considered a phenomenon that requires more than one level of description to be completely understood. Thus there are corpora that address various phenomena. However, there are few models for relating data from different levels of description to one each other.

Procedure

The definition of a tag on descriptive or procedural ground is almost never available, the theories behind the annotation cannot be looked at, tested or improved.

Context

The context dependency of tags remains opaque. Some properties are clearly context independent and can easily be predicted, some need complex detection procedures or are only tagged because of reasons of analogy. As a consequence, many corpora contain information which need not be learned or modelled by sta-

tistical applications. A clearer separation of these data is needed in order to be able to see what aspects must be further improved.

Ontology

Tags in corpora are not classified. Some of the tags in a corpus encode contextual information or physical properties of the speech signal, while others are linguistic labels. The status and validity of the information is not made explicit, but different types of data need different kind of further processing, part of the information will have to be updated or changed, which may be computationally expensive if all information is packed together.

Applicability

The application range of properties is not expressed in corpora. Some properties like *start* and [ç] are certainly intended to encode properties of one unit only. Information like the identity of a speaker, general voice quality etc. will be encoded once but may be important for large portions of the data. There is no data model for the placement and description of the applicability of types of information in corpora.

Dependency

Information units placed across a corpus are interrelated. A standard example is the property *start*, which has to be provided on various levels, e.g. for phones, words, sentences, utterances, and for many of these entities the value is the same. A change of values on one level implies the change of values on higher levels, but how and which is not made explicit in corpora. The same holds for more complex relations between theoretical entities.

Inflexibility

Corpus data are collections of static entries. The validity of the tags is limited by the validity of the theory that has been used to produce them. As most corpora are considered *final* once they are delivered in a set of files, any change in theoretical approaches cannot be applied to the corpus any more.

These aspects affect the use and reusability of annotated data and are partly responsible for the creation of more and more databases while existing corpora are not used any more without even having been fully exploited.

2. Aims

The critical evaluation of the status of speech databases above implies a model about what is needed in an ideal context of phonetic research and speech technology. These underlying general assumptions shall now be made explicit.

Understanding speech behaviour is the aim of phonetic research. Therefore, first of all models for entities, their properties, their mutual relationships and commu-

nicative functions need to be defined. These models must be able to segment and classify - recognise - elements of speech behaviour as well as they can be used to predict formally and acoustically adequate equivalents of meanings and intentions, i.e. synthetic utterances.

2.1 Use of corpora

If corpora are built to serve the above goal, then a central aspect of corpora is their use in the development and validation of theories. Two main purposes of tags can be distinguished:

Training: In many cases there are theories about the identity and the correct version of a label to be assigned to a given segment of speech, but no complete theory and equivalent algorithm has been found, yet. In this case tags are assigned by analogy. Sometimes, these tags are subject to further refinement. They can be used for the training of statistical approaches that serve to predict these tags with acceptable quality.¹ Corpora are needed to improve hypotheses and train models.

Development: For most levels of description one or more other levels of description are needed to determine the value of tags. Thus some corpora are used as test basis for additional levels of description whereas their correctness and completeness of the information is taken for granted. Corpora are needed to develop new theories.

Despite this distinction of tag types, most hypotheses and consequent linguistic tags cannot be considered final and may need revision.

2.2 Integration

Speech is not only a multi-layer event or a set of many separable streams of information the sum of which can explain linguistic functions. The level-wise description, segmentation and categorisation are abstractions which need to be reliable. Likewise, representation areas - acoustic, psycho-acoustic, physical, articulatory, and traditional linguistic descriptions - need to be integrated.

As mentioned above, a clear definition and distinction of the status of tags is missing in most corpora.

The standard notion of *corpus* is that of a collection of many utterances, ideally produced by many speakers, segmented and labelled on various levels of descrip-

¹ In this context it may not be forgotten that some of the units used for the description of speech are never questioned though they have not been proven to be most adequate for the description of speech, e.g. sounds and words.

tion, mostly linguistically motivated. Although the source of these data is often speech, signal aspects are rarely encoded symbolically. Apart from the amount of data to be represented, spectral properties of speech are mostly considered to be important for the decoding of segmental information and sentence modality only (fundamental frequency). Be it impracticable or not, signal properties are never represented symbolically in speech corpora.

Classical types of information in a corpus are circumstantial aspects and linguistic categorisation labels. Circumstantial aspects are aspects like date of recording, speakers, situation, recording personnel, hard- and software equipment etc. Linguistic categorisation labels are traditional entities like phonemes, words, sentences etc. and their categories.

Apart from the work in phonetics and speech technology, there are other data which are traditionally not conceptualised as speech corpora yet provide label information: Categorisation and recognition data from psycholinguistic experiments are used to test hypotheses. In most cases they are not further exploited, but provide important information on diverging perceptual and categorisation processes of speakers. Obviously, these corpora provide tags for small sets of linguistic stimuli, but many alternative labels per stimulus which also have to be predicted by a theories of speech.

An integration of different kinds of information would certainly be fruitful, as articulatory, acoustical, psycholinguistic and linguistic categorisation data can be evaluated and correlated. Although comparison and competition are essential aspects of progress in research, comparative evaluation of speech theories is impeded by different representations.

3. Ways

This section describes prerequisites and means to improve the accessibility to corpora.

3.1 Encoding

The first aspect to be addressed is the encoding itself.

Representation

A general representation format that is independent of the phenomenon or its theoretical description is required. An example of such a format is XML [2] which provides a general grammar specification formalism and is thus very flexible. Yet, encoding a corpus in XML only does not guarantee that other users of the corpus understand the meaning of the tags.

Relations

Ways of linking information, in order to express dependencies, simultaneity and other relations need to have separate and well described formal equivalents.

Definition

The representation of the meaning of a tag, its definition and the procedure or rules that selected the tag must be represented and retrievable when using a corpus. *Tag* denotes every symbolic entity in a corpus, and thus refers to the phenomena, property dimensions, values and relations that occur in the corpus. Although not a corpus itself, the formal grammar of the representation of these definitions must be the same as the one in which the speech corpus data have been encoded in order to be able to use the same software for their access.

Encoding

A general procedure for the encoding of theoretical description in databases needs to be defined. For the encoding of concepts and relations found, their property dimensions, values and the conceptual hierarchies between speech phenomena into the grammar – e.g. XML - the same procedures should be employed. This enhances the reuse as every person new to the corpus will know, how part-whole relations, inheritance of values, edges, and definitions can be found or recognised.

Especially the last two items is important for the user of the data as this is the kind of information that lets other people test the underlying theories and rules.

3.2 Corpus and systems

If corpora are used to improve theories then the long term goal of corpora must be that they become dispensable, i.e. because the theory for the description of language is complete (Figure 1).

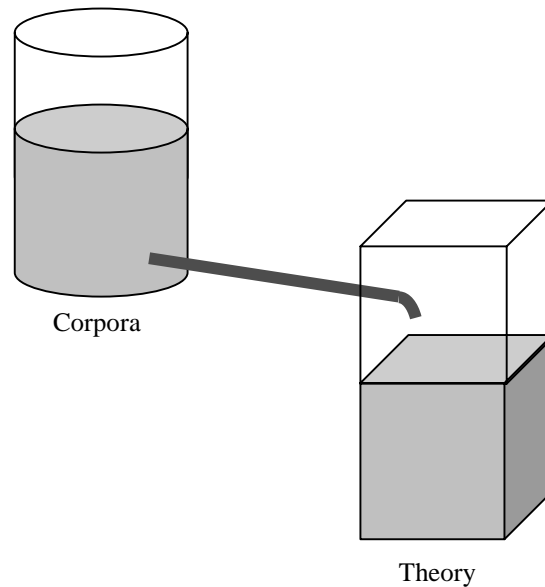


Figure 1. Corpora for the improvement of theories.

The measure for the completeness of a theory and the decreasing need for corpora or testing material is the correctness of predictions made by systems. In speech science one of the first levels on which this can be reached is the level of segmentation. Predictions can again be applied to segmentation and labelling tasks. Segmentation rules require properties of single (sub-)elements or sequences of (sub)elements as input. Classification rules require class-membership descriptions, specifications of inheritance of properties from other sources, or derivation procedures for properties that are calculated by using properties of other elements.

Predictions can be achieved by rules or by lexicon entries. The distinction of rules and lexicon entries should be seen as gradual: A *lexicon entry* is a rule which for a given element or sequence of elements provides a corresponding label. *Rules* are those entries that mostly require more abstract descriptions. Although the concept of *lexicon* is conventionally applied to words and associated information, the term can also be applied to other areas, e.g. that of prosody, where f0 configurations and corresponding utterance types are listed.

If it is the aim to model or predict information *A* by information *B* by having as few rules/entries as possible in a system, i.e. rules that are as general as possible, then this is a matter of the effectiveness, not of the correctness of a model. Effectiveness need not be the aim of theories as long as systems predict the same output. i.e. as many correct tags. The more abstract rules get, i.e. if they are defined in terms of abstract descriptions of properties, the more non-elementary representations and operations are needed to describe operations, thus more complex description models are necessary (Table 1).

rules systems	lexicon systems	require
less	more	entries
more	no	properties

Table 1. Comparison of rules and lexicon systems.

A corpus will always contain the actual state-of-the-art labels and categorisations. Existing rules are improved to predict segmentation and categorisation of speech data. Rules are available for each of the levels to be described and thus there will be interdependencies: Changing a theory or a rule will affect its output and thus the output of other rules as they might use the output of the former rule as input. Only if complex interactions of different hypotheses can be tested will it be possible to evaluate and compare the validity of theories. Thus there will be a network of interdependent aspects of speech levels (figure 2). Correctness of new rules and new hypothesis must be tested. Ideally, this is done by accessing older versions of the corpus.

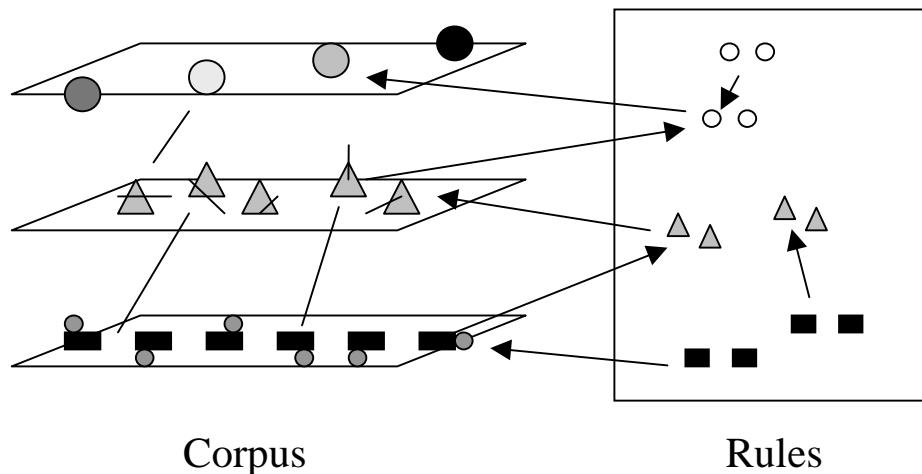


Figure 2. Interrelation of corpus and

3.3 Existing environment

Today, there is no such system as described above, and the aim presupposed is hard to achieve by individual research groups as it requires lots of data and expertise on the levels to be integrated.

Yet, some important steps have already been taken. In project MATE (European Union Telematics project LE4-8370), proposals for the effective encoding of dialogue data and tools for their support are developed. The encoding proposals provide discussion and strategies of interrelations between speech phenomena, theory and markup and give examples for five levels of description [3]. The features of the tool relevant for the subject here shall be briefly discussed. First of all the tool (*MATE workbench* [4]) supports the conversion of corpus formats into XML, reads XML data, stores them internally, provides (visual) support for the inspec-

tion and annotation of the data and saves changes or new annotation to XML files again. What makes this tool satisfy some of the requirements for research stated above are two other components, a query processor and a style sheet processor. The language of the query processor allows for evaluations, comparison, hierarchical evaluation, set operations of entities, and complex combinations of expressions by negation and logical operators. Additionally, special time operators for multilevel evaluation of speech events of various kind have been implemented [5]. The query processor can be used to search for complicated constellations of linguistic elements. The output of the processor is represented in XML again. Another description of the same fact would be: The query processor can be used to define phenomena. The output of the processor is a new corpus of annotation of tokens of this type of phenomenon. The output of a query does not only contain links [6] to the elements found, but also the query string. The style sheet processor – similar to XSL [7] - can be applied to this output and change or generate additional tags. As query language expressions are definitions and can thus be tags themselves, the described functionalities of the MATE workbench can be compared to the idea that procedural information of annotation can be represented and reproduced.

4 Conclusion

The lack of standardisation of corpus formats for speech data is obvious. Somewhat masked behind this syntactical phenomenon are other issues that impede access to speech corpora. Information and standardisation of the meaning of tags, the production history of markup, general guidelines for the conversion of theoretical, conceptual structures to syntactical markup equivalents, an integrated processing of information from different layers of description (phonetics, syntax, semantics) or phenomenological environments (acoustics, perception, function), and tools for the support of corpus annotation and theory work are important factors for improvements in speech research necessary to justify current efforts in corpus production. At least some of the critique and proposals of this paper have to be considered in order to develop theories and build systems that can be used to automatically tag (describe) and produce (predict) utterances and provide descriptions of their linguistic function and physical structure.

5 Acknowledgements

The work described here was founded by the European Union (Telematics project LE4-8370) and supported by IMS Stuttgart.

References

- [1] Klein, M., Bernsen, N.O., Davies, S., Dybkjær, L., Garrido, J., Kasch, H., Mengel, A., Pirrelli, V., Poesio, M., Quazza, S. and Soria, S.: *Supported Coding Schemes*. MATE Deliverable D1.1, July 1998.
- [2] W3C “*Extensible Markup Language*”: <http://www.w3.org/XML>.
- [3] Dybkjær, L., Bernsen, N.O., Dybkjær, H., McKelvie, D. and Mengel, A.: The MATE Markup Framework. MATE Deliverable D1.2, November 1998.
- [4] McKelvie, D., Isard, I., Mengel, A., Møller, Grosse, M., and Klein, M.: *The Mate workbench – an annotation tool for XML coded speech corpora*. To appear.
- [5] Mengel, A. and Heid, U.: *Query Language for Access to Speech Corpora*. Forum Acusticum, Berlin. (ASA, EAA, DEGA) 14-19 March 1999.
- [6] W3C, “*XML Pointer Language (Xpointer)*”, W3C Working Draft 9 July 1999. <http://www.w3.org/TR/WD-xptr>.
- [7] Clark, J. (editor) *XSL Transformations (XSLT), Version 1.0*, W3C Working Draft 9 July 1999 (<http://www.w3.org/TR/WD-xslt>).