

A Structure for Comprehensive Spoken Language Description

J. Bruce Millar

Australian National University
Canberra ACT 0200 AUSTRALIA
[bruce.millar@anu.edu.au]

Abstract

A comprehensive structure for the description of spoken language is presented. It includes static background data about the speech event, the representation of the speech event itself, and the value-added measurements that are made on the speech event. It extends the single-speaker and single microphone scenario to include complex speaking scenes. The structure is compared to other related schemes with the intent of illuminating potential upgrade towards a standard that can be widely accepted, fully extensible, and capable of being implemented on a wide range of information processing platforms.

Introduction

The decade of the 1990s has become distinguished for the emergence of corpus-based speech research and the wide use of spoken language data corpora by researchers other than those who designed, collected, and described these corpora. This situation poses a challenge to corpus developers to describe their corpora in a way that is highly intelligible for those who will use them and who will attempt to compare them with other corpora developed by other researchers.

Even simple speech events can benefit from being subjected to the rigour of this approach, but the major benefits are to be found in more complex speech scenes where the form of an utterance is strongly conditioned by its linguistic and situational precursors and by the ambience of its production. Indeed a mechanism to describe such complex speech scenes is an essential requirement if measurements of the speech are to be interpreted appropriately.

The current NSLD scheme of spoken language description (Millar, 1998) which arose from Millar (1992) and was utilised in earlier form in the ANDOSL corpus (Millar et al, 1994; 1998; Millar, 1996b) provides exemplars for this paper of a rich descriptive structure that can fulfil the need.

The Descriptive Tree

The structures described in this paper indicate one attempt to achieve a comprehensive framework for spoken language description that will enhance the intelligibility of such description. The essentials of this proposed framework are likened to the roots, the trunk, and the branches and leaves of a tree (Millar, 1994c and figure 1). The roots of the description are all those facets that pre-date the occurrence of the speech event but which influence the way in which it is produced. The trunk of

the description is the representation of the speech event itself as "seen" via a transducer, or a range of transducers, and as "stored" in an information processing system. The branches and leaves of the description are measurements or observations made on the speech event after it has occurred where the focus is on the perspective of the observer or the mechanism of the measurement.

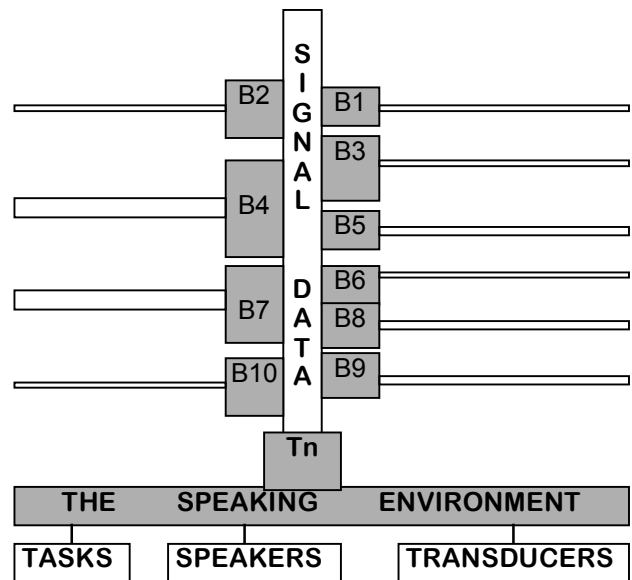


Figure 1: Description Tree for One Transducer and having ten branches of post-hoc descriptors.

The Roots

Millar (1992) proposed ten classes of description for spoken language data. A subset of those classes can be seen to form the pre-conditions of the speech event whose acoustic form is the data we need to describe. These classes are the nature of the speaker(s), the nature of the speaking task(s), the nature of the transducer(s), and the nature of the speaking environment in which speakers and transducers are located (figure 2).

The Speaker The nature of the speaker has many dimensions not all of which are available for measurement. Indeed some significant facets of a given

speaker can at best be estimated from one or more related measures. Ideally we would want to know the size, shape, surface conditions, muscular strength of the vocal apparatus which will provide the scope of the speech sounds, both static and dynamic, that the speaker can utter. Further we would want to know the linguistic, social, and phonological status of the speaker which will provide certain constraints on the way we expect the speaker to speak. The NSLD scheme provides a wide range of direct and indirect types of such measures which relate to the nature of the speaker (Millar, 1992; 1996b; 1998).

The Speaking Task The nature of the speaking task also has many dimensions which comprise all the reasons why the speaker chooses to utter speech on the occasion of the speech event. In many data corpora used in speech research today the reasons are fairly well constrained by a scientifically designed data collection procedure. The need for the parameters of the task to be carefully described is crucial especially as less structured speaking scenes are accessed. These parameters may include initial instructions given to the speaker, but also the reaction of the speaker to those instructions can influence the nature of the speech which is elicited. Most speaking tasks are driven by "prompts" which may be words to the read, visual images to react to using speech, or even roles to adopt in an interactive situation.

The Transducer The nature of the recording transducers including their optional settings, their orientation with respect to the speaker(s), and their proximity to any reflective surfaces can all be relevant to the form of the signal that they capture. Where non-acoustic transducers are involved other parameters such illumination (for video) or precise placement (for physiological) may be needed to provide a full description of the collected data.

The Speaking Environment The nature of the speaking environment includes both passive and active audio, and maybe video, influences and the position of the speaker, or speakers, with respect to recording transducers. Some static aspects of the environment may be deduced from the recorded signal but there is no substitute for "ground truth" when separating out the causes of signal characteristics.

Summary of Roots So with these three bundles of roots of the descriptive tree we can set the scene for the analysis of the speech prior to the speech event itself. The speech event grows out of these roots when the impetus to speak occurs.

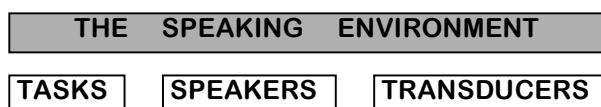


Figure 2. The Root Components of the Description

The Trunk

The speech event itself is probably the simplest aspect of spoken language to describe. Each transducer in the "speech field", including microphones, cameras, physiological transducers, et cetera, generate a signal data stream which is converted into digital form and stored in a signal data file. This trunk of the descriptive tree comprises all the signal data files produced plus a data description file (DDF) for each one (figure 3). The DDF provides information which links the data file to a specific transducer and to a specific speech event by means of a transducer number, whose position and characteristics are defined in the "root" section, and the start time/date of the event. Information is also given about the location of the data so that it can be readily accessed., data conversion and/or transformation characteristics, and the format in which it is stored in the data file.

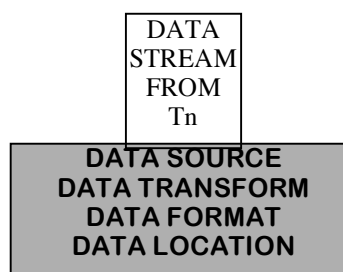


Figure 3. The Trunk Components for One Transducer.

The Branches and Leaves

The branches and leaves of the description are the post-hoc descriptors which add value to the data by recourse to external judgements and/or signal processing methods. These descriptors are amongst the most difficult to define in a reasonably general way. They fall into two broad categories: those which are the result of potentially complex signal processing which itself has many parameters, and those which are the result of human judgement using a symbolic "transcription" scheme and relying on training and experience of the operator. The former include such tasks as "formant tracking" where informative derived parameters are deduced following complex processing which relies on careful setting of parameters of the processing. The latter include all manner of phonetic and linguistic annotation and have been the source of much discussion in many places including the corpora and labelling working group of the COCODA organisation. They may also include the description of anomalies that have been observed in the data recording or processing. The essential feature of the branch is that it transforms the data that is in the trunk into the value-added data that is in the leaves (figure 4). The leaves, representing the surface form of the derived analysis parameters or the annotations, have meaning only if their mode of derivation (the branch) from the speech event (the trunk) is clearly described.

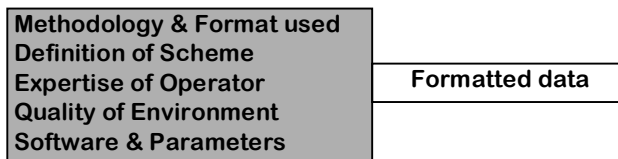


Figure 4. The Branch Component for One Description.

Summary

The three partitions of description afforded by the description tree analogy allow us to picture the comprehensive descriptive structure required to fully describe a spoken language event.

Roots for complex environments

The vast majority of spoken language research that has been conducted to date has utilised data that has been collected serially from single speakers using a single transducer. The strong interest in expanding this simple situation to study dialogue (e.g. Anderson et al., 1991) and speech in groups (e.g. Crawford et al., 1994) and further to look at multi-modal analysis of spoken language “scenes” injects a strong challenge for existing structures for spoken language description. In this section we look at the structural changes in descriptive schemes needed to accommodate such developments.

Single Speaker and Single Transducer (SSST) Environments

The SSST environment is the simplest situation where all aspects of the speaking environment (ENV) can be efficiently combined in a single environment category. This category will include the passive characteristics of the speaking space (ENV-1), the active acoustic environment (ENV-2), the type of transducer (ENV-3), the channel between transducer and recording medium (ENV-4), and the spatial relationship between the speaker and the transducer (ENV-5). These subsets are summarised in Table 1. Each of the five subsets if this total description of the speaking environment will have multiple parameters as exemplified in Millar (1998).

Component	Content
ENV-1	Passive Acoustic Characteristics
ENV-2	Active Acoustic Components
ENV-3	Transducer Characteristics
ENV-4	Channel Characteristics
ENV-5	Speaker-Transducer Relationship

Table 1: NSLD Components of an SSST Environment.

Single Speaker and Multiple Transducer (SSMT) Environments

The SSMT environment differs from the SSST environment by the duplication of the latter three parts of the total ENV description. It may also introduce some

new ENV sub-categories which are specific to the additional transducers used. This is clearly the case when a camera is introduced as in that situation the active and passive aspects of illumination of the space become important. Therefore the first two ENV sub-categories need to be duplicated according to the number of transduction modes in use and the latter three according to the number of transducers (Table 2). Further the introduction of such additional transducers requires that we define a spatial origin against which to locate transducers, speakers and other active sources of acoustics or illumination.

Component	Content
ENV-1a/b	Passive “modal” Characteristics
ENV-2a/b	Active “modal” Components
ENV-3n	Transducer “n” Characteristics
ENV-4n	Channel “n” Characteristics
ENV-5n	Speaker-Transducer Relationships

Table 2: NSLD Components of an SSMT Environment.

Multiple Speaker and Single Transducer (MSST) Environments

The MSST environment is one that it often encountered when a planned SSST environment suffers interruption from an unplanned source. Whether by accident or design, moving from an SSST to an MSST environment involves the duplication of ENV-5 only. However, in addition to environmental variables there are likely variants in speaking task across different speakers (TSK: and SPK: in Millar (1998)). This can be clearly seen in such simple multiple speaker situations as the Edinburgh “MAP” task, where the task of one speaker is to question the other, and the task of the other is to respond to those questions.

Component	Content
TSK_n	“Nth” task description module
SPK_n	“Nth” speaker description

Table 3: NSLD Additions for an SSMT Environment.

Multiple Speaker and Multiple Transducer (MSMT) Environments

The MSMT environment represents the most complex of speaking scenes with static components (Millar, 1994b). This may for example involve the description of data collected from a panel discussion which comprises individual or strategically placed microphones and one or more video angles of the group and maybe close-up shots of individuals. A system to automatically understand the spoken language interactions encoded in such data will benefit from a full description of the active and passive components of the speaking space (ENV-1/2-Audio/Video), the details of all transducers and their channel processing (ENV-3/4-all), and of course all the speaker-transducer spatial relationships (ENV-5-all).

Dynamic Speaking Scenes

The next stage of description allows for movement of speakers and transducers in the course of the recording. This will require a time parameter to be added to the positional components of ENV-2, ENV-3, and ENV-5 which will allow the spatial relationship descriptors to be updated at certain intervals.

Managing a complex Trunk

The trunk of the description of a complex environment will simply be multiple parallel data streams. These may well have different sampling rates and different formats but each data stream will have all such descriptors defined in its transducer-specific data description file (“Tn” in figure.1).

Attaching branches and leaves

The branches and leaves of the description tree comprise multiple value-added derivations from the recorded speech event. They will typically comprise “analysis” branches resulting from signal processing, “annotation” branches resulting from either pattern recognition or expert evaluation, “anomaly” branches resulting from careful observation of procedures, and “acknowledgement” branches applied by administrators.

Multiple Annotations

The concept of multiple annotations of a speech recording is well established. The International Phonetic Association (IPA) has sanctioned “at least two levels of transcription” for the segmental annotation of an utterance of any language (IPA, 1989). The value of more than two levels has been promoted and cited widely (e.g. Barry and Fourcin, 1992; Tillman and Pompino-Marschall, 1993). Although certain annotation schemes are widely used, there are many local variants of such schemes in use too, so, for complete transparency, a branch descriptor module, which completely defines the process by which the leaves on the branch are produced, is required. The component parts of the branch descriptors for annotation data are outlined in Table 4 and are further detailed in Millar (1996a).

Component	Content
LAB-1	Annotation Environment Parameters
LAB-2	Annotation Data Presentation
LAB-3	Annotation Scheme Detail
LAB-4	Annotation Strategy and Method
LAB-5	Annotator Training and Experience

Table 4: NSLD Components of Annotation Description.

The relationship between the physical quality of signal data (LAB-2 in table 4) and the ability of a human annotator to effectively apply a given annotation scheme will clearly depend on the level of annotation being attempted. Some international agreement on the salient physical parameters against which this relationship may be assessed has been reached (Millar, 1994a). No further progress on the nature of this relationship has been reported.

Similarly, there is no formal agreement on the way in which to describe annotation schemes (LAB-3 in table 4),

although proposals for action to relate all “local schemes” to a universal standard such as that proposed by the IPA (Esling and Gaylord, 1993).

Multiple analyses

In most cases it is required to perform a variety of analyses on the speech signal data in order to support further description of the data. These may, for instance, include an excitation frequency detector, a formant tracker, or a low-order cepstral coefficient processor. Such analysis is used to provide evidence for prosodic labelling, segmental phonetic labelling or an automatic segmental labeller respectively. In the subset of cases where it is valuable to retain these intermediate forms of the data a suitable structure needs to be provided. The component parts of the branch header are outlined in Table 5.

Component	Content
ANA-1	Analysis Algorithm Source
ANA-2	Analysis Parameters Applied
ANA-3	Analysis Data Format

Table 5: NSLD Components of Analysis Description.

Multiple anomalies

A keen observer of any process of spoken language data collection will note instances, or even persistent influences that are not a part of the explicit design. These observations will range from glitches in equipment performance to the attitude or attention to task of the speakers. A scheme to code such observations when they are offered can result in additional post-hoc data which can influence the way that the performance of speech technology systems built upon these data may be interpreted (table 6). Such anomaly observations may be multiple in that they can be noted by different observers at different stages of processing - during production, processing, or annotation.

Component	Content
ANO-1	Speaker Performance
ANO-2	Recording Environment
ANO-3	Linguistic Influence
ANO-4	Voice Quality
ANO-5	Technical Problems
ANO-6	Incomplete Material

Table 6: NSLD Components of Anomaly Description.

Multiple acknowledgements

In any large scale spoken language data corpus many people will contribute their intellectual property to various stages of the development of the corpus. There may also be restrictions on data use resulting from contracts with the speakers, with a funding body, or with any contributor to the corpus. The attachment of acknowledgement and/or restriction statements to the machine-readable data is one way to alert users of the corpus of their responsibilities in these matters.

Summary

The latest version of NSLD has incorporated structures that allow the range of complexity described above to be represented. This of necessity can make the NSLD system look overly complicated when presented to potential users. However it should be emphasised that while NSLD aims for comprehensive scope, it can be applied in highly reduced form when there is no need for the complexity of which it is capable.

Comparison with other schemes

The value of file headers or associated description files has long been realised as necessary aids to the interpretation of a wide range of possible digital representation of speech data. More recently the recognition of the diversity of such interpretive aids has encouraged the development of "standard" descriptive systems. Our modular sets of NSLD descriptors are intended as a contribution to this search for acceptable standards. It is likely that any emerging standard will include some of the best features of these pre-standard offerings. In this section we examine NSLD in comparison with other contending systems.

Schemes in widespread use

Two schemes for machine-readable spoken language description in common use are the extensible header structure of the American National Institute for Standards in Technology (NIST), and the associated description file of the European Speech Assessment Methodologies (SAM) project.

NIST/SPHERE - header structure NIST's SPHERE system of file headers (NIST/SPHERE Standards, 1994) and associated software management have gained widespread acceptance through the DARPA sponsored competitive speech technology research programme and the data brokerage of the Linguistic Data Consortium. This structure was also adopted by the Australian National Database Of Spoken Language (ANDOSL) project for a simple file header standard. However this header structure in no way replaces the application of NSLD to the ANDOSL data and is indeed automatically derived as an NSLD subset. The modular object-oriented structure of NSLD allows rich description without the inefficiencies of repetition that would be required if its richness were replicated in a header structure similar to that used by NIST.

SAM - Associated Description Files The SAM project's associated description files (SAM Standards, 1994) and wide-ranging supporting software have had a strong influence on the development of the NSLD proposal. In its present form, the SAM associated description file lacks explicit modularity and clear modes of extensibility. NSLD uses the basic SAM key and value tuple but explicitly groups them into object-oriented modules and defines the means to link additional options at root, trunk, and branch levels. The basic key/value tuple of SAM and NSLD could usefully be expanded to use the NIST key/format/value scheme.

Annotation Schemes

The goal of a "standard" in machine-readable phonetic annotation has a long history. Conflicting philosophies and methodologies have ensured a delay in reaching this goal. In NSLD the form of annotation scheme used for the attachment of linguistic labels to a speech event is simply represented as a parameter in the module that attaches any branch to the trunk. NSLD assumes multiple annotations will be present and that they maybe one variety of phonetic (IPA-CRIL; SAMPA; MRPA, WorldBet, etc), prosodic (SAMPROSA; ToBi), orthographic, discourse, or any other.

Data Structures

Object-Orientated Design The NSLD structure is object-orientated in its design in that it aims for redundancy-free yet complete representation of any spoken language scene. It has not however at this stage been coded within an environment that explicitly supports object-orientated structures or processing. Contemporary simpler systems have illustrated how this may be done using both the Lisp and Prolog languages respectively (Karjalainen and Altosaar, 1993; Draxler, Tillmann and Eisen, 1993). The use of an advanced computer language system to encode the full richness of NSLD is a future task.

Hierarchically Related Annotation A extension to the present model is the use of a linked set of branches where separate levels of annotation are linked by a set of rules. The QuickSig system (Karjalainen and Altosaar, 1993) treats the total phonetic space as a hierarchy based on parent/child relationships within an object-oriented programming environment. This system creates persistent phonetic annotation data in a single data file. The EMU system (Harrington and Cassidy, 1997) provides for a hierarchy of levels of labelling within which timing at lower levels can propagate to higher levels. Each labelling level is stored as a separate file and in addition a heirarchical label file which stores the relationships between the individual label files. An NSLD extension of this general kind would require a generic branch descriptor covering all levels of labelling in the hierarchy as they are interactively related.

Formal Structure The NSLD scheme currently exists as a formatted ASCII tabulated list comprising key-value tuples segregated into object classes It shares a strong similarity of overall objective with the Text Encoding Initiative (TEI) as described by Burnard (1995) which is formally structured using SGML as a basis. Whereas the domain of the TEI is the multimedia "text", the domain of NSLD is the "spoken language event". Both initiatives are designed to provided a rich descriptive structure that addresses the challenges of reusability, seamless integration, and loss-free interchange of valued data. It is clear that the structures of the TEI could be applied to the branch structures of NSLD and may even be more widely applicable.

The format of the spoken language description is less important than its content, its coverage, and its acceptance by practitioners. It must at the same time look simple to those who wish to describe simple events, and supportive

of rich complexity to those who have complex events to describe. It seems likely that the proposals of both the NSLD and the TEI will suffer from a similar “image problem”, that of overkill for simple situations. The simple formatted text file implementation used for the ANDOSL project (Millar, Harrington and Vonwiller, 1998) stands as an example representing simple speaking scenes. The value of more formal implementation for complex scenes has yet to be demonstrated but to the likely benefits of re-use, integration and interchange, mentioned above, must be added convenient interface to relevant database management systems.

Conclusions

This paper has drawn together many short presentations which have as their aim the extension of the NSLD scheme of spoken language description. The simple analogy of a tree has facilitated the integration of many facets of spoken language description both conceptually and graphically. Methods of implementation have been touched on lightly in this paper but are likely to be the major focus of further work in this area when considerations of validation and management of data from complex speaking scenes will be centre stage..

Acknowledgements

I acknowledge the support and encouragement of many members of the Australian Speech Science and Technology Association and especially those involved in the ANDOSL project. I am grateful for the feedback generated by several presentations of parts of this material to COCOSDA workshops. I also acknowledge the expert assistance of Arthur McGuffin in providing computer programming support to this project over several years.

References

- Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H., Weinert, R. (1991) The HCRC Map Task Corpus, *Language and Speech*, 34, 351-366.
- Barry, W.J., Fourcin, A.J. (1992) Levels of Labelling, *Computer Speech and Language*, 6, 1-14.
- Burnard, L. (1995) Text Encoding for Information Interchange, <http://www.uic.edu/orgs/tei/info/teij31>
- Crawford, M., Brown, G.J., Cooke, M., Green, P.D. (1994) Design, Collection and Analysis of a Multi-Simultaneous-Speaker Corpus, *Proc. Institute of Acoustics*, 16(5), 183-190.
- Draxler, C., Tillmann, H.G., Eisen, B. (1993) Prolog Tools for Accessing the PHONDAT Database of Spoken German, *Proceedings of the 3rd European Conference on Speech Communication*, (pp.191-194) Berlin, Germany, 21-23 September.
- Esling, J.H., Gaylord, H. (1993) Computer Codes for Phonetic Symbols, *Journal of International Phonetic Association*, 23, 83-97.
- Harrington, J.M., Cassidy, S. (1997) *Techniques in Speech Acoustics*. Kluwer Academic Publishers.
- International Phonetic Association (1989) The IPA Kiel Convention Workgroup 9 Report: Computer Coding of IPA Symbols and Computer Representation of Individual Languages, *Journal of International Phonetic Association*, 19, 81-82.
- Karjalainen, M., Altosaar, T. (1993) An Object-Oriented Database for Speech Processing, *Proceedings of the 3rd European Conference on Speech Communication*, (pp.183-186) Berlin, Germany, 21-23 September.
- Millar, J.B. (1992) The Description of Spoken Language, *Proc. 4th Australian International Conference on Speech Science and Technology*, (pp.80-85) Brisbane, Australia, 1-3 December.
- Millar, J.B. (1994a) Relationship between Physical Quality and Transcription, *Notes from COCOSDA workshop 94*, (pp.32-33) Yokohama, Japan, 22-23 September.
- Millar, J.B. (1994b) Multi-Speaker and Multi-Sensor Description, *Notes from COCOSDA workshop 94*, (pp.34-38) Yokohama, Japan, 22-23 September.
- Millar, J.B. (1994c) Taxonomy and Infrastructure for labelling, *Notes from COCOSDA workshop 94*, (pp.39-40) Yokohama, Japan, 22-23 September.
- Millar, J.B., Vonwiller, J.P., Harrington, J.M., Dermody, P.J. (1994) The Australian National Database of Spoken Language, *Proc. ICASSP-94*, (pp.97-100) Adelaide, Australia, 19-22 April.
- Millar, J.B. (1995) Overview of Australian Spoken Language Corpus Resources and Standards, *Notes from COCOSDA Workshop 95*, (pp.8-21) Madrid, Spain, 22 September 1995.
- Millar, J.B. (1996a) Labelling the Labler, presented to the *COCOSDA Workshop 96*, Philadelphia, USA, 7 October, URL=http://cslab.anu.edu.au/~bruce/cococal/cococal97/presentations/labelling_the_labeller.
- Millar, J.B. (1996b) National Spoken Language Description scheme, URL= http://andosl.anu.edu.au/andosl/general_info/cd_doc/mark2/docs/all_nsld.pro
- Millar, J.B. (1998) *NSLD-98 Definition Profile*, URL= <http://cslab.anu.edu.au/~bruce/NSLD>
- Millar, J.B., Harrington, J.M., Vonwiller, J.P. (1998) Data Resources for Australian Speech Technology, *Journal of Electrical and Electronic Engineers Australia* (in press).
- NIST/SPHERE Standards (1995) URL=<http://www.icp.grenet.fr/Relator/standnist.html>.
- SAM Standards (1994) URL= <http://www.icp.grenet.fr/Relator/standsam.html>.
- Tillmann, H.G., Pompino-Marschall, B. (1993) Theoretical Principles concerning Segmentation, Labelling Strategies and Levels of Categorical Annotation for Spoken Language Database Systems, *Proceedings of the 3rd European Conference on Speech Communication*, (pp.1691-1694) Berlin, Germany, 21-23 September.