Universität Hamburg

Sonderforschungsbereich Mehrsprachigkeit

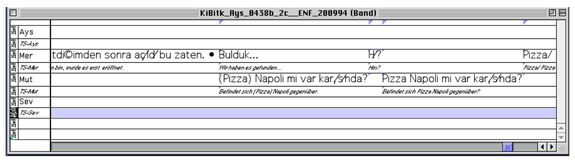
Deutsche Forschungsgemeinschaft DFG

EXMARaLDA

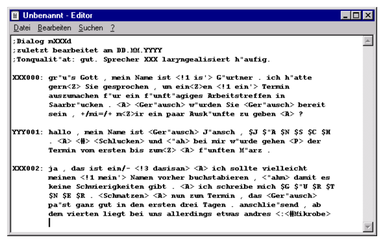
EXtensible MARkup Language for Discourse Annotation

LE BABYLONE DES FORMATS

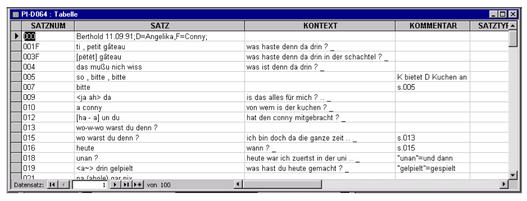
„Particular bodies of data are created with particular needs in mind, using formats and tools tailored to those needs, based on the resources and practices of the community involved. Once created, a linguistic database may subsequently be used for a variety of unforeseen purposes, both inside and outside the community that created it.“ [Bird/Liberman 2001*: A formal framework for linguistic annotation*]



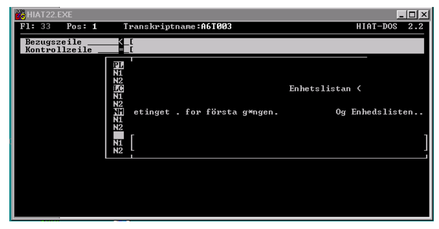
syncWriter



Verbmobil



dBase /Access



HIAT-DOS

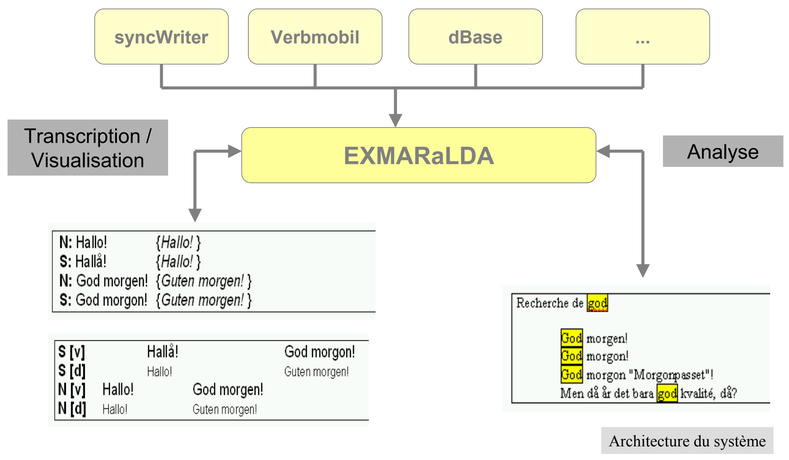
La plupart des treize projets du SFB "Mehrsprachigkeit" (Centre de recherche sur le multiliguisme) travaillent avec des transcriptions de langue parlée. Ces donnés sont très diverses à beaucoup d'égards:

- les **langues analysées**: allemand, français, portugais, italien, espagnol, anglais, turc, japonais, basque, russe, chinois, luganda, suédois, danois, norvégien, …
- le **cadre théorique**: recherche phonétique, syntaxique, analyse de conversation, …
- la **forme des donnés**: les phénomènes transcrits, leurs représentations graphiques, les étiquetages grammaticaux etc.
- le **type d'interaction** enregistrée: interactions entre enfants et adultes, émissions de radio, conversations interprétées, discours de présentation, conversation "libre" entre plusieurs personnes, interviews etc.
- les **moyens informatiques**: outils de transcription (syncWriter, HIAT-DOS, dBASE, éditeur de texte, cf. ci-dessus), systèmes d'exploitation (Windows, MAC OS, Linux)

Bien que la diversité qui s'exprime dans les quatre premiers de ces points soit certainement souhaitable et productive - et d'ailleurs inévitable - d'un point de vue théorique, la diversité qui résulte du dernier point pose des problèmes pratiques graves, cf. la citation ci-dessus.
Les problèmes décrits dans cette citation apparaissent aussi dans le travail quotidien du SFB: il n'est pas possible d'échanger facilement les données entre les différents projets ou d'élargir une transcription donnée par des annotations supplémentaires, parce que chaque projet choisit pour ses données des outils informatiques et des formats de mémoration qui ne sont appropriés que pour le but particulier que poursuit ce projet.
Le but du projet "Base de données multilingue" serait alors de réunir les données linguistiques des différents projets sous une forme généralisée permettant un échange facile, un élargissement flexible et une mémoration à longue terme. Le résultat concret de ce travail sera un corpus de la langue parlée multilingue contenant des transcriptions de plus de 1500 heures d'enregistrements.
Le système EXMARaLDA est la composante centrale pour la construction de ce corpus.

UNE SOURCE, PLUSIEURS CIBLES

“The new opportunities are not yet being fully recognized and exploited by linguists [...] Texts are still seen as objects in book format, with words running in horizontal lines from left to right. Annotations are added to these horizontal lines. But book format is an attribute not of speech, but of Western writing systems. There is no reason beyond established custom and practice to present speech in this way. On the contrary, since there are often several annotations relating to the same piece of data, book format is in many cases inappropriate. The use of book format without consideration of other possibilities is based on a confusion between the organization of the data itself, and the presentation of the data on the printed page.” [Knowles 1995: *Converting a corpus into a relational database*]

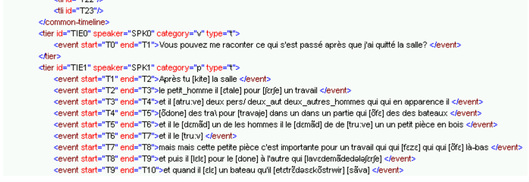


Architecture du système

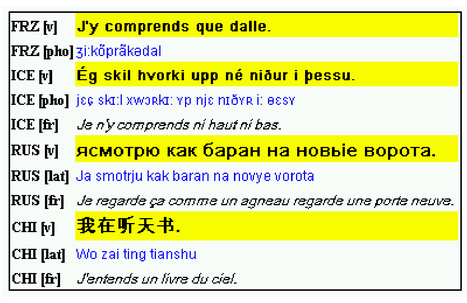
EXMARaLDA est un système de transcription computationelle basée sur l'idée des "**Annotation graphs**" (graphes d'annotation) de Steven Bird et Mark Liberman (Bird/Liberman 2001). Comme telle, EXMARaLDA définit un format pour mémoriser des transcriptions sous une forme indépendante:

- d'une *langue* particulière ou – plus exactement – du système d'écriture utilisé pour la transcription. Comme EXMARaLDA met en œuvre la technologie **UNICODE**, il est possible, par exemple, d'utiliser l'alphabet latin, l'alphabet cyrillique et l'alphabet phonétique international (A.P.I.) dans un même document.
- d'un *logiciel* particulier. Le format de mémoration dont EXMARaLDA se sert est **XML** (Extensible Markup Language), un standard qui est compris par les systèmes d'exploitation et les langues de programmation les plus courants. Ainsi, il est facile de construire de nouveaux outils pour des données existantes, dans le cas où les outils qui font partie du système même ne sont pas appropriés à un certain but.
- d'un *cadre théorique* particulier (dans le sens décrit ci-dessus): EXMARaLDA est basé sur un modèle abstrait d'une transcription de discours. En **séparant le contenu logique** d'une transcription, c'est à dire les unités structurales utilisées (comme les mots, les énoncés etc.) et leurs relations (temporelles et linguistiques), **de sa forme graphique**, c'est à dire de sa représentation sur une page ou sur un écran, il est possible de réduire dans une large mesure la diversité de différents systèmes de transcription mentionnée ci-dessus. Ainsi, une seule transcription peut avoir différentes formes de visualisation, et il est possible de mettre dans un seul document les différents niveaux de description linguistique, même si ceux-ci semblent incompatibles à première vue.

Avec cette architecture de système il est possible d‘avoir plusieurs modes d‘édition et de visualisation pour un même document, et d‘analyser les données diverses dans un cadre commun.



XML

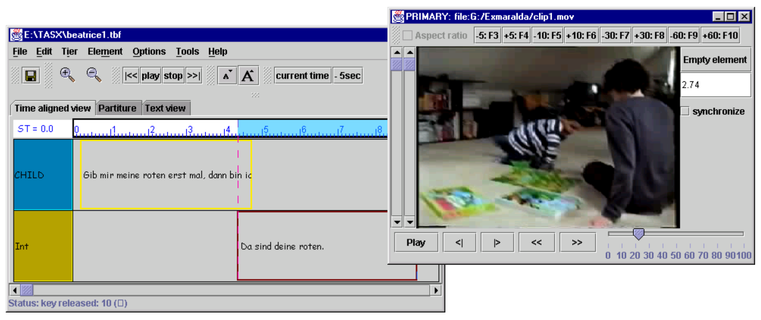


UNICODE

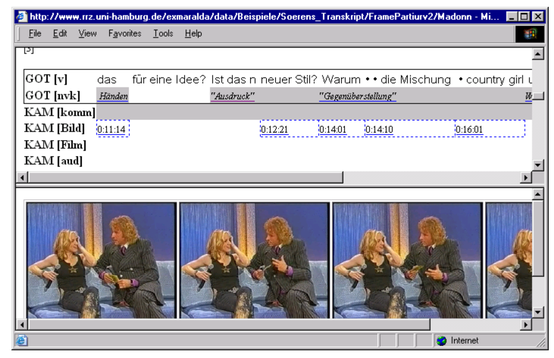
OUTILS DE TRANSCRIPTION

“[…] a code which is convenient for a machine, which may be essential for a machine, may be extremely awkward for humans. […] there will have to be software interpreters provided, that will allow us to move between human and machine without any expenditure of effort whatever. I can’t see any alternative to this, and I doubt if anyone who is managing a large spoken resource would want to dispute it. Humans who are going to use transcribed text will want human conventions as much as the transcribers. They will want to see in front of them something that is reasonably readable.” [Sinclair 1995: From *Theory to Practice*]

EXMARaLDA Partitur-Editor



TASX Annotator



Visualisation en HTML

Thomas Schmidt | thomas.schmidt@uni-hamburg.de | http://www.rrz.uni-hamburg.de/exmaralda