The ELRA Newsletter

January - March 2004

Vol.9 n. 1

Contents

Signed articles represent the views of their authors and do not necessarily reflect the position of the Editors, or the official policy of the ELRA Board/ELDA staff.

ISSN: 1026-8200

ELRA/ELDA
CEO: Khalid Choukri
55-57, rue Brillat Savarin
75013 Paris - France
Tel: (33) 1 43 13 33 33
Fax: (33) 1 43 13 33 30
E-mail: choukri@elda.fr
Web sites:
http://www.elra.info or
http://www.elda.fr

Dear Colleagues,

Our special issue to honour the memory of Antonio Zampolli was well appreciated by our readers. Many of you welcomed this tribute paid to our friend and colleague. We hope that all enjoyed reading Antonio Zampolli special paper for his Institute and the contributions from his friends. This newsletter is the 1st for year 2004; we will briefly elaborate in this introduction on the strategies which need to be adopted in cooperation with other data centres, companies, and R&D labs at the international level for the promotion of LRs and of HLT more generally. In parallel, we will try to offer an overview of the actions carried out by ELRA and ELDA. During the last few months, the need for new LRs was clearly expressed by people involved in the development of new technologies and research projects through user needs surveys conducted by ELRA and other partners, within e.g. the Nemlar (Network of Euro-Mediterranean Language Resources) and ENABLER (European National Activities for Basic Language Resources) projects.

Indeed, to better adjust its services offered to the HLT community, ELDA has been strengthening its LRs production and collection activities, in particular through its participation in European and French projects.

Those projects involving the launch of new LRs, mainly SLRs, include, at the European level, C-Oral-Rom, with the production of speech corpora in 4 Romance languages (Spanish, Italian, French and Portuguese); TC-Star, a project focusing on speech-to-speech translation for which ELDA has been involved in the LRs identification i.e. spoken multilingual aligned corpora; CHIL, aiming at improving interactivity between human and computers, for which ELDA collected and annotated speech data; CLEF, the Cross-Language Evaluation Forum, with the production of a French written corpus which will then be used by the participants in the evaluation campaign; and OrienTel, for the launch of multilingual interactive services in Mediterranean countries, where ELDA has supervised the production of a speech database in Jordan. At the French level, projects for which ELDA created new LRs include Neologos, with the production of new SLRs, as well as Lexitec and Euradic, with the production of mono- or bilingual lexicons and specialised dictionaries covering the English, Spanish, German, French, Arabic and Greek languages These 3 projects have been conducted within the action line dedicated to the LRs production under the Technolangue programme.

Another area strongly connected to the LRs production and collection is the LRs validation. To illustrate the importance of LRs validation, we may highlight ELRA's involvement in the area since early 2001: ELRA has set up a network of technical centres to take care of the validation of the SLRs and WLRs presented in its catalogue. Doing so, ELRA ensures the quality of the LRs it distributes and offers the HLT community a better visibility. This work is achieved in cooperation with SPEX (SPeech Expertise center), ELRA's Validation Centre for SLRs, located in the Netherlands, and CST (Center for SprogTeknologi) in Denmark, which heads ELRA's VCs for Written Language Resources. The validation of LRs has been further promoted with the launch a few years ago of a Bug Report Service. This service allows the user of a LR purchased from ELDA to report the imperfections he/she may find. It is currently available only for SLRs but should be implemented in the near future for WLRs. The Bug Report Service can be accessed via the ELRA web site, www.elra.info. The HLT evaluation has become over the past years a milestone activity in the field, allowing developers to assess the performances of their systems and offering the community comparative results. ELDA participates in European projects where HLT evaluation is central, namely CLEF (Cross-Language Evaluation Forum), CHIL (Computers in the Human Interaction Loop), and TC-Star (Technology and Corpora for Speech to Speech Translation). The evaluation activity is also prominent at the French level, through the Technolangue programme: ELDA is the coordinator of the EVALDA platform, which includes 8 evaluation campaigns focusing on new technologies for the processing of the spoken and written French language. Some of the campaigns are still open to new participants: if you are interested, please contact the ELDA team to obtain more information.

If you would like to learn more about ELRA and ELDA's activities, and about the European and French projects both bodies are involved in, you are kindly invited to visit our web sites, at www.elda.fr and www.elra.info. At the beginning of 2004, ELRA and ELDA have been strongly involved in the preparation of the 4th edition of the Language Resources and Evaluation Conference. LREC 2004 took place in Lisbon, from 24th to 30th May. The next ELRA newsletter, which should be distributed during the summer time, will give you an overview of the conference, with some sessions' and workshops' summaries. For the time being, you will find here 2 papers on very different topics: the first one presents ARABTALK®, a text-to-speech synthesis system for Arabic language; the second article is about specialized content and terminology. As usual, the new resources added in our catalogue can be found in the last pages. This is the last newsletter under the Presidency of Joseph Mariani, who participated in the ELRA adventure in close relationship with Antonio Zampolli from the very early days of the Relator project, and who succeeded him as second ELRA president in 2002. He will be replaced by Bente Maegaard, who was elected ELRA president at the General Assembly during LREC'2004 in Lisbon.

Joseph Mariani, President

Khalid Choukri, CEO

ARABTALK® An Implementation for Arabic Text To Speech System

Yasser Hifny, Shady Qurany, Salah Hamid, Mohsen Rashwan, Muhammad Atiyya, Ahmid Raghed, Galaal Khallaaf

This paper describes the ARABTALK® Text-To-Speech (TTS) synthesis system, developed at RDI, for Arabic language. ARABTALK® is a state-of-the-art corpus-based concatenative TTS system. The system employs Artificial Neural Networks (ANN) statistical prosody-based models for duration, energy, and global pitch contour prediction. In addition, it has a real time synthesis by selection algorithm to explore large speech corpus. ARABTALK® has a Hidden Markov Model (HMMs) based procedure to automatically time-align new voices transcriptions to their acoustic phoneme boundaries. In this framework, a mature phonology framework has been developed and many perfect rule-based models were utilized in the process of letter to sound conversion. The system is multi-user and safe-threaded enabled for server based applications. This research aims to advance the process of developing high quality Arabic TTS synthesis, which yields natural and human sounding Arabic voices.

Introduction

Corpus-based unit-selection concatenative text to speech paradigms are the state-of-the-art high quality natural TTS systems. ARABTALK® is one of these systems, which is developed specially for Arabic language. This paper describes the overall architecture, several components of the system, and linguistic concepts for Arabic. Many components of the system are corpus based like statistical prosody models and corpus preparation.

This paper is structured as follows. Section 2 describes the Arabic phonology developed to generate phrases targets and phonological features, which are utilized in the prosody prediction and unit selection process. The approach to prepare the speech-aligned corpus is presented in section 3. The description of the statistical prosody models is presented briefly in section 4. Section 5, describes the two kinds of units used for concatenation, monophone and diphone. Section 6, describes the unit selection process. Finally, section 7 summarizes our conclusions and the expected future work.

A framework for Arabic phonology

ARABTALK® has a mature Arabic phonology framework. Many problems have to be defined and solved in order to achieve automatic letter to sound conversion and provide all the necessary information for other components of the system like phonological to acoustic components mapping (duration, energy, and intonation models) and unit selection process.

Our vision for Arabic language suggested the following tasks to be solved in order to have a reasonable output:

Standard Arabic language has twenty-eight consonants and six vowels. The six vowels are divided into three short vowels and three long vowels. The long vowels have similar spectral properties like their short vowel version with longer durations than the short vowel version. However, the current system has 41 phonetic letters by adding extra phonemes to consider the effect of the pharyngealized phonemes.

Morphological diacritics are the diacritics of a word characterized by word structure and it is one of the core tasks in order to have automatic letter to sound conversion. Arabic orthography does not consider short vowels within the word structure. RDI has a statistical solution developed by the NLP group to predict possible short vowel patterns for a sequence of words [1].

Syntactic diacritics are the short vowels assigned to the end of each word and they are assigned on the basis of syntactic analysis for the whole phrase. The prosody generation and the unit selection algorithms are affected directly by syntactic diacritics as the actual databases are recorded in a natural way. In order to avoid developing syntactic Arabic analyzer, we suggested and introduced a novel corpus-based approach as a workaround to predict the syntactic diacritics based on HMM Tagging methods. This approach will be developed by NLP group and integrated to the system in the future versions. Currently, we assign a blind default diacritic type for the syntactic diacritics during the automatic letter to sound conversion. They could also be supplied manually.

Consonant clusters are eliminated as Arabic has a prosodic nature to remove heavy pronunciation. The consonant clusters are three adjacent consonants, which may result during the physical pronunciation, and are eliminated by inserting a short vowel between the first and second consonant. The type of the short vowel is selected by using simple rule based model.

Phonetic grammar validation is a procedure to ensure that a given phrase could be parsed correctly by the syllabification algorithm where an Arabic syllable must start with only one consonant and the syllabic structure prevents three consonants or two vowels to appear adjacently. This problem usually happens when mixing an Arabic text and a non Arabic text (written in Arabic orthography) in one sentence.

Letter to sound conversion for Arabic usually has simple one to one mapping between orthography and phonetic transcription for given correct diacritics. Some simple rule-based methods are used to complement the generation of the phonetic transcription.

Syllabification for Arabic language as Arabic has only six syllable types (CV, CVC, CVV, CVVC, CVCC and CVVCC). The last three types usually appear at the end of a phrase only due to their heavy pronunciation. The durations of the consonants and the vowels within these three types are known to be longer than the other remaining

types. The number of vowels and the number of syllables in an Arabic phrase must be equal. Hence, any stream of valid Arabic syllables could be accurately parsed according to these rules.

Morphological stress assignment is described as predictably falling on a particular location in the word, depending on the internal structure of the syllables making up the word [2]. So, Arabic stress is known directly from the word syllable structure of a word. Arabic stress assignment is different from English language, which uses the stress as a free phoneme. Hence, Arabic stress is a morphological stress and not a lexical stress. The stress patterns are derived from an implementation of the stress assignment procedure, which is a combination of the work that has been developed by the phoneticians [3, 4, and 5]. Further enhancements will be integrated to the current model when we develop Part Of Speech (POS) tagger for Arabic.

Currently, the system does not have any automatic procedure to assign different accents degrees to word sequence. The accent degree for a word could be assigned manually or could be ignored during the transcription process. The last word of a phrase has a higher accent degree by default in the current implementation. Moreover, an algorithm that changes the accent degree for a sequence of words is implemented. The primary objective of this procedure is to assign different accent degrees for function words and content words. Data driven approaches for prosodic phrasing and accent label predictions will be integrated in the next versions, as we have suggested and developed the specifications for a new general-purpose text corpus for Arabic "AL-KHALIL" [6]. This corpus will have rich annotation tags for syntactic, prosodic phrasing, and accents. These tags are assigned to guide statistical models to discover some rules about Arabic grammar and Arabic semantics. Parsing a given utterance results in a prosodic tree, which is constructed to represent the different levels of the phonological description and the relationship between these levels. The output of this linguistic component is utilized by prosody models and unit selection process.

Database preparation

The system has two databases one for male speaker at 22 kHz sampling rate (one hour) and the other for a female speaker at 16 kHz sampling rate (four hours). The speech is coded into 12 dimensional MFCCs plus log energy and their derivatives. The EGG signal is recorded with each utterance to support pitch synchronous analysis and prosodic modification if necessary during the synthesis process. We use HMMs based Viterbi alignment procedure, developed at RDI for this purpose [7].The Viterbi alignment procedure can be summarized as a problem of searching time boundaries for known sequence of HMM models for phonemes. Since the best state sequence, which is known to be the Viterbi path, is obtained during decoding process, time boundaries can be obtained directly. This process is illustrated in figure 1.

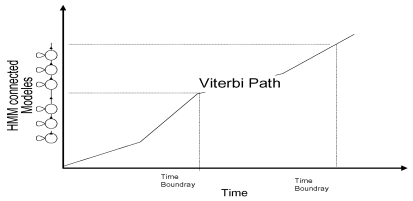

Figure (1) Viterbi Alignment

Actually, Viterbi (forced) alignment procedure results are reasonably well but the labels need to be more accurate for a synthesis database than for recognition. Hence, we did many manual corrections and we have developed many tools to correct and move boundaries for similar error patterns automatically.

Prosody modeling

As shown in figure 2, Phonology to prosody modeling is achieved via BP neural networks. The system utilizes three different neural networks to estimate the duration for each unit, the average energy per sample for a unit, and the global intonation contour for each phrase. The authors described the duration model of the system and its prediction accuracy in details [8]. The global intonation contours used for training were extracted from the speech and each syllable was represented by eight values from the contour. The predicted phrase contour is a smoothing version of the concatenation of the predicted syllable pitch contours. During the unit selection the target costs are weighted scores between the predicted duration/pitch and the extracted values of duration/pitch for a unit. The training and testing procedures are based on the NN simulator that is developed for similar task [9].

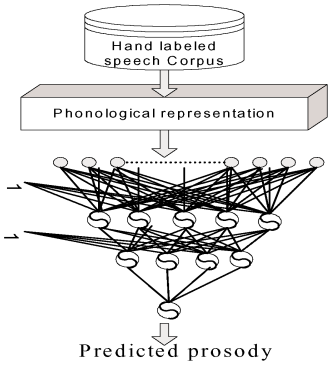

Figure 2: Phonology to Prosody Mapping

Database Unit

Current version of ARABTALK® is built so that the concatenation unit can be either monophone or diphone. In monophone case, the system was very sensitive to segmentation errors which degraded the intelligibility of the produced speech. Using human revision for the automatic segmentation - though tedious and time consuming- limited this problem although it wasn't completely solved.

As a solution for the system sensitivity to segmentation problems, the monophones were replaced by diphones as concatenation units. The diphone unit segmentation was made using the automatically segmented monophones and the boundaries of the diphone were taken from half of the first phoneme to half of the second phoneme. Segmentation can be done in two more ways: a) starting and ending at nearest pitch marks. b) Automatic segmentation of diphone units using HMM. This is left for future improvements.

The use of diphone units as concatenation units mainly solved the problem

of system sensitivity to segmentation problems and the generated speech was much smoother at the concatenation points. This increased the system intelligibility. The direct use of automatically segmented speech was also made possible.
For the context clustering of the diphone units, the same tree structure was used as the monophone case. The only difference is that the first part of the diphone was considered as the previous phoneme and the second part was the next phoneme.

Unit selection

Unit selection algorithms are developed to explore large databases in order to minimize prosodic and spectral modifications for high quality speech synthesis [10]. They aim to select the best sequence of units that match the required targets from a speaker database by Dynamic Programming (DP). The selection process is based on a combination of target cost and continuity cost. Target costs measure how a unit in the database matches a target unit in the target phrase. The continuity cost is a distortion measure for coupling two neighboring units. In general, the unit selection algorithms are similar for the problem of searching the best state sequence in the HMMs using the Viterbi algorithm. The transition probabilities and observation scoring have the same role of the target and continuity costs in searching the best state sequence.
The search space (the searched lattice) is considered large both in the horizontal and vertical directions in Arabic language because of two different factors. The horizontal direction, which is defined by the number of target units, is relatively larger than in the English language since the phonetic letters per word are approximately doubled after adding the short vowels to each word. The vertical direction, which is defined by the available number of unit candidates, is also large. The actual number of vowels in a database is huge because the Arabic language has only six different vowels and every syllable must have a vowel in the target phrase. ARABTALK® implementation of the unit selection process is optimized to achieve a real time performance. As shown in figure 3, the current implementation utilizes *Candidate Caching*, which reduces the search space, and *Continuity Caching*, which reduces continuity cost calculations.

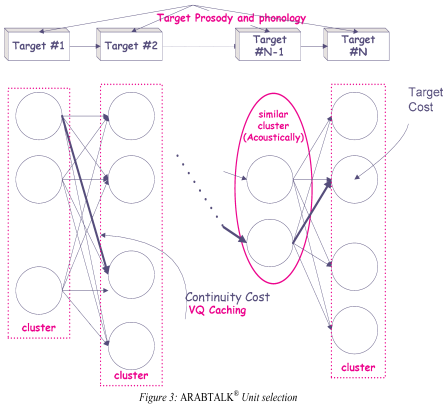

Figure 3: ARABTALK® *Unit selection*

Figure 3: ARABTALK® Unit Selection

Continuity Caching is achieved by Vector Quantizing (VQ) the spectral features (MFCC) of the coupling frames (first and the last frames) for each unit. The quantizer is based on Principle Component Analysis (PCA). In the current implementation the number of the centroids is 1024. A distance matrix between the centroids is saved and used during the synthesis process as a look up table to approximate the continuity costs.
During the synthesis process, ARABTALK® target costs imply only weighted prosodic cost between target pitches and durations with respect to candidates' values. The continuity costs are differently weighted at the coupling boundaries for syllables and words. These boundaries are defined while generating the targets for a phrase. The acoustic costs are not considered in the current implementation since the cluster units are very similar acoustically.

Phonology Level	Feature Description	Feature count	Possible range
Phoneme	Sound Type	11	1 to 13
	Voicing Type	11	1 to 5
	Consonant Type	11	1 to 9
	Type Of Articulation	11	1 to 13
	Place Of Articulation	11	0 to 15
	PhonemeID	11	0 to 41
	Fuzzy Emphatic	11	0 to 1
	Emphatic Type	11	0 to 1
	Shadda	11	0 to 1
	Tanween	11	0 to 1
Syllable	Phoneme Position	1	1 to 4
	Count of Phonemes	1	2 to 4
	Accent Degree	1	0 to 4
Foot	Syllable Position	1	1 to 10
	Count of Syllables	1	1 to 10
Phrase	Foot Position	1	0 to 3

Table 1: Clustering Questions Description

Candidate Caching aims to cluster the candidates of the same units [11]. Hence, the online search uses the units of the selected clusters to build the DP lattice. The clustering process is achieved by decision trees and only spectral similarity measure is used while splitting process to evaluate unit similarity. In this work, detailed
WAGON clustering program, output trees loader was available from EST tools [12].

Summary

The overall architecture and general features of ARABTALK® Text-To-Speech system for Arabic language has been presented. An online demo is available at "www.rdi-eg. com/rdi/research/arabtalk.asp". The system is corpus-based and has many statistical models. It has real time unit selection with different caching methods. Our current research has many directions to improve the quality of the output speech. For example, different basic units

will be investigated as we search better smooth continuity between the selected units. An automatic data reduction procedure, which offers flexibility in the database size, will be integrated in the next version for the handheld applications. Towards better intonation contours, the group will investigate methods based on ToBI labeling methods. Finally, parametric synthesis like H+N methods may be developed in order to have better coupling methods between units.

Acknowledgements

The authors wish to express their thanks to members of RDI speech department for their support in constructing databases. We thank the NLP group who made the necessary modifications for their work to match our vision and to be integrated with the current system. We thank Wael Hamza, who developed the first speech synthesis system at RDI. We would like to thank Alan Black, Wael Hamza, Muhammad Afify and Christof Traber for their fruitful discussions. Many thanks to Christof Traber who supplied us his PhD thesis hard copy.

References

[1]Muhammad Atiyya, "A large-scale computational processor of the Arabic morphology", MSc thesis, Cairo University, Egypt, 2000.

[2]Kenneth de Jong and bushra Adnan, "Stress, duration, and intonation in Arabic word-level prosody", Journal of phonetic,Vol. 27,3-22, 1999.

[3]Ibraheem Anis, "The Sounds of Language", Dar Al Nahda Al `rabia Press, Cairo, Egypt, 1961.

[4]Tammam Hassan, " Research Methods in Language", Al Risala Press, Cairo, Egypt, 1955.

[5]Salman h. al-ani, "Arabic Phonology: An Acoustical and Physiological Investigation". The Hague, Netherlands: Mouton and Co., 1970. "Janua Linguarum" series practica 61. Translated into Arabic, 1983.

[6]Ahmid Raghib, Yasser Hifny, Mohsen Rashwan, "ALKHALIL, General purpose Arabic text corpus for the applications of Text To Speech synthesis", Technical report, Research and Development International (RDI) company, Cairo, Egypt, 2001.

[7]Wael Hamza and Mohsen Rashwan, "Concatenative Arabic speech synthesis using large database", In Proceedings of ICSLP2000, vol. 2, pages 182-185, Beijing, China. 2000.

[8]Yasser Hifny, Mohsen Rashwan, "Duration Modeling for Arabic Text to Speech Synthesis", In Proceedings of ICSLP2002, pages 1773-1776, Denver, Colorado, USA, 2002.

[9]Yasser Hifny, "Online Arabic Handwriting Character Recogntion ", MSc thesis, Cairo University, Egypt, 2000.

[10]A. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using a large speech database". In ICASSP-96, volume 1, pages 373--376, Atlanta, Georgia, 1996.

[11]Black, A, and Taylor, P. "Automatically clustering similar units for unit selection in speech synthesis", Eurospeech97, Rhodes, Greece, 1997.

[12]Paul Taylor, Richard Caley, Alan W. Black, Simon King, "Edinburgh Speech

Muhammad Atiya
RDI The Engineering company for computer systems development
23 Omar ibn elkattabst. Mohandeseen
PO Box 406 Imbaba
EG Giza
Egypt
m_atteya@rdi-eg.com
Technical Arabic NLP team leader
Tel + 20 10 102 59 63
Fax + 20 2 33 82 166

Semantic Interoperability and Language Resources

Christian Galinski

Most users are interested not in the (hardware and software) tools, but in content. There are many kinds of content, including specialized content (representing domain specific knowledge in some way or other - including terminology). Terminology in specialized content is in most cases embedded in or combined with other kinds of content (mostly specialized texts). In order to make content development less expensive (because of its labour-intensiveness), we need new methods of content creation (and the respective workflow management): **net-based, distributed, cooperative creation of structured content**.

In principle **all** e-Content items/units (which under a comprehensive content management perspective are or should be based on a metadata approach and on unified data modelling principles and requirements) should be prepared and maintained in such a way that they fulfil the requirements of

- single-sourcing => uninhibited re-usability

- resource-sharing => (net-based distributed) cooperative content development

- universal accessibility => incl. access by persons with special needs.

This gives **interoperability** a new dimension - the fundamental requirement for achieving the aims of the **Semantic Web**. ISO/TC 37 is gradually moving into this area, bringing in its competence and experience with the data modelling of terminological data and other language resources from the point of view of “content” complementary to the point of view of the ICT approaches.

Definitions

Content in technical terms is defined as
- text (textual data, incl. all kinds of alpha-numeric data),
- sound (audio data),
- image (graphical data),
- video (multimedia data).

However, from a "semantic" point of view, this is completely insufficient. Under a mobile content (mContent) perspective, today, content - including terminology - is from the outset:
- multilingual,
- multimodal,
- multimedia.
and should be prepared in such a way that it meets **multi-channel** requirements.

Content should also be prepared in such a way that it is **re-usable** in all kinds of applications, especially the e-...s, such as:
. e-learning,
. e-government,
. e-health,
. e-business,
. etc.

Sociolinguistics distinguishes between **general purpose language** GPL (or common language in the generic sense) and the **special purpose language** SPL (or specialised language in the generic sense). One of the main characteristics of SPL is its high share of terminological units, which are indispensable for:
- domain (or professional or subject-field) communication,
- representation of specialised (i.e. subject-field related) knowledge,
- access to specialised (i.e. subject-field related) information.
In this context we speak of the "specialised languages" (SPLs) of the various subject-field/domain expert communities, which agree on their linguistic conventions (mostly geared towards the written form of their respective SPL) not necessarily in conformance with GPL conventions. Furthermore, quite some SPLs comprise - at least in their written form - many (and many different types of) non-linguistic representations, which also belong to content.

Content seen as content items

To a large degree e-Content - especially domain specific content - takes the form of **textual data** (i.e. alphanumeric data of a textual nature), which, from a formal point of view, are composed of **language resources** (LRs , including text corpora, speech corpora, grammar models, lexicographical and terminological data).
Concepts in terminology are corresponding to (material and immaterial) objects in the real world (which comprises also human society and culture). Concepts are mental constructs functioning as 'first order representation', whereas the corresponding terms (or other kinds of concept representation) have the role of 'second order representations'. Concepts have the function to condense information and to provide a certain order for the 'things' around us. This order in its totality is in a constant state of change, depending on knowledge change and also on the point of view taken by the observer. This order at any given point in time not only works at the level of concepts, but has implications on higher levels of scientific-technical theory building.

Under the aspect of semantic interoperability, which is indispensable, if present e-Content and future mContent (comprising multilingual content in eBusiness, eLearning, eGovernment, eHealth and all other e-…s) shall really be utilized efficiently and effectively (taking into account content management), one soon recognises that there are different types of 'mental constructs', which can be called concepts in a wider sense. In terminology itself there are different types of terminologies (based on different types of concepts), which can be subsumed under the respective concept systems such as:
- logical concept systems (which can be hierarchical, non-hierarchical or hybrid),
- ontological concept systems (which also can be hierarchical, non-hierarchical or hybrid),
- other kinds of concept systems (which again can be hierarchical, non-hierarchical or hybrid),
or which can be typologised as
- regular scientific-technical terminologies (tending towards a hierarchical type of concept system),
- social-science and humanities oriented terminologies (tending towards a network type of concept system),
- nomenclature-type of terminology (following specific naming rules for naming the nomenclature classes),
- other.

In addition there are conceptual units, which can be called 'terminology phraseology', which often serve as a prestage in the terminologisation of linguistic units to become terms (representing a distinct concept). Vice versa there are terminological units, which are de-terminologised and become lexical units of the general purpose language (GPL).

In general GPL, too, there are different types of"'mental constructs" usually called meaning. There are words and their morphological components, as well as collocations etc. There is a natural process of "terminologisation" of GPL units into terminology as well as "de-terminologisation" of terminological units into GPL usage. Brain research proves that there is no clear borderline between scientific-technical categorizing and classifying thinking and GPL communication, where meanings of words and utterances show a high degree of ambiguity. But exactly because of that they are highly productive in coping with any communicative situation.

All this is quite "object-oriented", as concepts correspond to objects. Every object - whether material or immaterial - is part of the whole universe and, therefore, is ultimately related to all information of the universe (which cannot - for the shear volume of this information - be processed by the human brain). Conceptual thinking - a "condition humaine" of mankind - is absolutely necessary for the human brain to condense information and reduce information volume in such a way that it can be made instrumental for coping with everyday life.

Given this immense volume of specialized information (i.e. scientific-technical or professional information), one or more meta-levels of condensation are necessary: documentation languages (i.e. indexing and retrieval languages) like classification schemes

and thesauri. They are needed for several purposes, among others:
- subdividing volumes of information into “manageable” portions,
- indexing of information for re-use,
- retrieval of indexed information,
- browsing in information,
- etc.

If there are many such documentation languages for different purposes, one further meta-level becomes necessary: umbrella classification schemes. For the sake of data processing of such documentation languages the respective metadata, datamodels and metamodels have to be defined.

In product description and classification (PDC), more “object-related” data are needed for each product (which can also be a service). Some types of products' designations still belong to the traditional domain of terminology. But what about series, models (and sub-models) and components as well as (mass-produced or individually produced) products? Here names of products or identifiers or barcodes can become synonyms. Some of the additional data to distinguish series, models, components, individual products, names (of makers, distributors, …), etc. can be used as attributes, others as 'traditional' properties and characteristics. Among others the relatively new field “ontology” in data modelling tries to find solutions to structure this mass of information. However, only such methodological approaches are viable, which produce results that are “reproducible” under same or similar conditions.

A simple overview on terminology usage thus shows a terrific mess in naming and defining elements such as class, attribute, property, characteristic, dictionary, etc. This mess calls for a clarification of basic concepts in eBusiness, etc. in order to make content fully interoperable (including re-usability, single sourcing and resource-sharing under an extended content management perspective) across all kinds of applications. If terminology belonging to same or similar 'objects' remains as fuzzy as it is today, the various expert communities for metadata approaches, ontology, eLearning, content management, documentation, and last but not least terminology cannot communicate properly with each other. They would conceive competing and even contradicting methodological approaches in order to cope with their respective problems. The very basic requirements of content management (in its broadest meaning), such as single-sourcing (in order to achieve optimal re-usability) and resource-sharing (in order to save human efforts in content development) cannot be met in this case.

Within the framework of the Workshop CEN/ISSS/eCAT "Multilingual electronic catalogues and product classification" of the (Information Society Standardization System of the European Committee for Standardization) an attempt is made to clarify some or most of the conflicting terms concerning product description and classification so that communication across subject-fields becomes possible and terminologists can find their role in formulating basic principles and requirements for multilingual content development. Who else would have the know-how and competence to do this?

Standardization

In principle, **all** e-Content items/units (which under a comprehensive content management perspective, are or should be based on a metadata approach and on unified data modelling principles and requirements) should be prepared and maintained in such a way that they fulfil the requirements of :
- single-sourcing resulting in uninhibited re-usability,
- resource-sharing as a basis for (net-based distributed) cooperative content development,
- universal accessibility (incl. access by persons with special needs).

For the sake of a comprehensive re-usability (under a broad content management perspective) we need more methodology standards than what exists today. Such methodology standards can be sub-divided into:

. Standardisation - Top-down including:
- harmonisation of metadata,
- unification of principles and methods of data modelling,
- standardisation of meta-models,
- standardisation of workflow methodology

. Standardisation - Bottom-up including (for instance in e-business) :
- product classification - terminologies,
- product identification,
- ontologies,
- e-catalogue data,
- LRs.

By using net-based distributed cooperative working methods on the basis of methodology standards, some of which do not yet exist, content development will become much less expensive in the future than today through extensive net-based **co-operation** on the basis of **standards**.

In the field of terminology standardization, ISO/TC 37SC 1, SC 2 and SC 3 take care of the standardization of terminological principles and methods as well as of certain terminological applications. The individual terminologies - as far as they are needed for the work of other TCs in ISO, IEC and other standards bodies - are standardized by those TCs. LR related principles, methods and certain applications are standardized by ISO/TC 37/SC 4 "Language resource management"; which was established in close cooperation with ELRA (and in particular with the pro-active support of Antonio Zampolli). ISO/TC 37/SC 1 also takes care of the terminology of terminology science, terminology applications and language resource management. So there is a quite comprehensive framework for standardization activities in the field of terminology and other language resources in place.

The Semantic Web is conceived as the **global e-Content infrastructure** for:
- e-Business, e-Learning, e-Health, e-Government, e-Health, and other e…s,

and - if it shall be efficient and effective,
- must **provide rules and procedures as well as organizational frameworks** to guarantee or at least support **different kinds of interoperability**, such as

technical, operational, syntactic and semantic interoperability:
. throughout the enterprise/organization,
. between enterprises/organization,
. within industry consortia,
. between industry consortia (urgently needs open standards),
. among different e…s,
. between different language communities,
and also **within the world of standards** (which also needs further development and harmonization).

World-wide content updating and maintenance mechanisms

The results of this e-Content related unification, standardization and harmonization efforts need to be regularly and constantly updated/maintained according to developments in science and technology, and even more so to the expectations on the user side. Furthermore, in the age of the Semantic Web, computers have to communicate in seemingly natural language, which - contrary to true natural language - has to be more or less **unambiguous**. The developing information society, therefore, will need many repositories of :
- certain types of data: authority data, attributes, values, etc.,
- terminological data of all sorts,
- names (of countries, currencies, organizations, etc.),
- non-linguistic representations of knowledge,
- certain data elements, metadata, data categories, etc.,
- data structures/datamodels & metamodels,
- interchange formats, XML schemas,
- syntactic communication protocols, messages, etc.,
- interfaces,
- data dictionaries,
- typologies, taxonomies, nomenclatures, ontologies, etc.,
- etc.

supplementing existing ones . This will require a systematic approach to the establishment of :
. maintenance agencies - whenever there is a need for a high degree of authority and high stability over time,
. registration authorities - securing a high degree of consistency over time and more or less strict registration rules,
. registries for codes, words (and word elements, terms, term elements, etc.) and for attributes, values, etc.,
which have to take care of these repositories in a distributed, but well coordinated way. This calls for a policy of the standardization system, how to deal with such **maintenance agencies, registration authorities and data registries.**

Given the need for many more (and different types of) maintenance agencies, registration authorities and registries, it needs a coherent framework for:
- the 'objects' to be taken care of by these MAs, RAs and repositories,
- the degree of authoritativeness of each type of object,
- the objectives of standardized and non-standardized updating/maintenance procedures,
- the terms of reference of these MAs, RAs and repositories,
- the work methodology as well as workflow management methods to be used in the updating/maintenance process,
- etc.

Such a policy for a distributed, however well coordinated framework for all kinds of content items today only exists in a rudimentary form. The development may well end up in a **network of distributed (federated) MAs, RAs and registries** becoming **the backbone of the e-content infrastructures** of the Semantic Web. Given the requirement for coherence of the objects taken care of in these MAs, RAs and repositories, the standards bodies will find new opportunities for standardization activities; but they will also have the societal responsibility to take the lead.

Copyright for terminological data and other kinds of textual content

Concerning the content of the abovementioned MAs, RAs and Registries, there is a copyright problem. According to ISO/TC 37 standards, a terminological entry consists of one (or more) entry term(s) (or abbreviation, symbol, etc.) and a definition. The term represents the underlying concept in a short, 'symbolic' form, whereas the definition represents the characteristics of the concept in a “descriptive” form. If terminology is about representing concepts, then non-linguistic representations - be it graphical or other symbols or be it complex formulas or other kinds of non-linguistic representation of the characteristics of the concept in question - can equally represent concepts (and have to be acknowledged side by side with terms). In fact, as a result of technological development, the ways and means of concept representation are increasing, while the share of non-linguistic representations of concepts (and other kind of knowledge). There are also other kinds of IPRs on non-linguistic representations than on terms and definitions being textual data.

The definition (acc. to ISO 10241) must not be a complete sentence, because term and definition have the relation of an equation (which means that they should be exchangeable in any given occurrence in a text). This does not support their copyright. Words (even multi-word terms) cannot fall under copyright. The minimum constituent element of a text, which can fall under copyright is a sentence (generally speaking). This is a “forma”' assessment. The main question of copyright and IPRs in general is, however, whether the idea which is expressed is “original”, whether it constitutes a “work”. But the definition of a terminological entry only reveals, what is state-of-the-art of scientific-technical development (i.e. which is correct not 'true' - and therefore common knowledge - at a given stage of development). So it cannot be 'original', even if experts have spent a lot of time on the formulation of the terminological entry. In any case, the use of individual entries falls under “fair use”, especially if one cites and acknowledges it properly. The copyright statement in dictionaries (and other printed works) do not conform to law, if they try to impose stricter provisions than the law itself. (Only

if there is a well specified bilateral contract between two parties, stricter provisions can be implemented acc. to civil law).

There is, however, the EU Directive for the protection of databases (or substantial parts hereof). If one extracts a substantial number of entries from a dictionary, one should try to obtain the written permission of the publisher. If one wants to avoid that in the case of minor extracts, one could send a letter telling that one will use so-and-so-many entries of his/her publication considering this as fair use. In addition one could state that the proper citation of the source would result in publicity for the book thus increasing its commercial value.

Scientific/academic ethics should 'morally' prohibit to deprecate definitions in order to circumvent copyright, but strict enforcement of (this not really enforceable) copyright would inevitably lead to this undesired consequence.

Similar considerations have to be made with respect to the contents of all kinds of MAs, RAs and Registries (for all kinds of repositories).

Outlook

The "keep it simple, stupid!" principle in data modelling invariably results in very high costs (usually at the users' expense, who do not get what they actually need - but more often they cannot specify their needs). Given the complexity of the semantic interoperability requirements to be observed already today, experts from various quarters, such as :
- terminology and other language resources (incl. the multilinguality and multimodality aspects),
- internationalisation and localisation (incl. cultural diversity and psychological aspects),
- information design (incl. accessibility aspects),
should take the initiative and prepare fundamental basic standards cutting across all application fields with respect to multilinguality, multimodality, cultural diversity and related issues (covering also to some extent general cultural diversity, psychological and accessibility aspects). The application specific communities have to develop standards with the basic principles and requirements of the respective application field. New professional profiles for content development will have to be designed and implemented at educational institutions to provide the market with content developers able to cooperate with system designers and maintenance experts in developing also the most appropriate data models and metamodels conforming to - hopefully - international standards.

The standards bodies not only will find new opportunities for standardization activities (and new business opportunities through related services), but also have the societal responsibility to develop **a network of distributed (federated) MAs, RAs and registries** becoming the **backbone of the e-content infrastructures** of the Semantic Web in order to secure the consistency and coherence of all objects (i.e. content items and other objects) taken care of in these MAs, RAs and Registries all across the Semantic Web.

Christian Galinski
TermNet - International Network for Terminology
Aichholzgasse 6/12
A-1120 Vienna
Phone: +43 1 817 44 99
Fax: +43 1 817 44 99 44

NEW RESOURCES

ELRA-W0015 Le Monde Text Corpus

Year 2003 of Le Monde Text Corpus is now available in .XML format.

Price for ELRA members per year of data (research use only)	240.91 euro
Price for non members per year of data (research use only)	313.18 euro

ELRA-W0036/04 Le Monde Diplomatique Text Corpus in Arabic

Electronic archiving of "Le Monde Diplomatique" articles in Arabic from 1998. The corpus is available in an ASCII text format.
French and English versions also available.

Price for ELRA members per year of data (research use only)	46 euro
Price for non members per year of data (research use only)	69 euro

ELRA-S0162 Hempel

This corpus contains 25.5 hours of recordings by 3,909 German speakers with a total of 184,240 spoken words, made via public phone lines (fixed network only). The contents are free monologues answering the question: "Was haben Sie in der letzten Stunde gemacht?" (What did you do within the last hour?). The database is conformant with the SpeechDat Exchange Format.

	ELRA members	Non-members
For research use	755 Euro	1,010 Euro
For commercial use	4,755 Euro	5,010 Euro

ELRA-S0158 Turkish OrienTel Database

The Turkish OrienTel database comprises 1700 Turkish speakers (921 males, 779 females) recorded over the Turkish fixed and mobile telephone network. The OrienTel database has been collected by the Orta Dogu Teknik Üniversitesi, Ankara, Turkey.
This database is partitioned into 1 DVD. The speech databases made within the OrienTel project were validated by SPEX, the Netherlands, to assess their compliance with the OrienTel format and content specifications.

Speech samples are stored as sequences of 8-bit 8 kHz A-law. Each prompted utterance is stored in a separate file. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information.

Each speaker uttered the following items:

1 isolated single digit, 1 sequence of 10 isolated digits, 5 connected digits (1 prompt sheet number of 6 digits, 1 telephone number of 6-15 digits, 1 credit card number of 14-16 digits, 1 PIN code of 6 digits, 1 spontaneous phone number), 1 currency money amount, 2 natural numbers, 3 dates (1 spontaneous e.g. date or year of birth, 1 prompted date, 1 relative or general date expression), 2 time phrases (1 time of day spontaneous, 1 time phrase in word style), 3 spelled words (1 spontaneous e.g. own forename, 1 city name, 1 real word for coverage), 5 directory assistance utterances (1 spontaneous e.g. own forename, 1 city of childhood, 1 frequent city name, 1 frequent company name, 1 common forename and surname), 2 yes/no questions (1 predominantly ”yes” question, 1 predominantly ”no” question), 6 application keywords/keyphrases, 1 word spotting phrase using embedded application words, 4 phonetically rich words, 9 phonetically rich sentences.

The following age distribution has been obtained: 982 speakers are between 16 and 30, 431 speakers are between 31 and 45, 274 speakers are between 46 and 60; the age of 13 speakers is unknown.
A pronunciation lexicon with a phonemic transcription in SAMPA is also included.

	ELRA members	Non-members
For research use	40,000 euro	45,000 euro
For commercial use	47,500 euro	57,000 euro

ELRA-S0159 German spoken by Turkish OrienTel Database

The German spoken by Turkish OrienTel database comprises 332 Turkish speakers who spoke German (167 males, 165 females) recorded over the German fixed and mobile telephone network. The OrienTel database has been collected by the Bavarian Archive for Speech Signals at the Institut für Phonetik und Sprachliche Kommunikation at Ludwig-Maximilian University in Munich, Germany. This database is partitioned into 1 DVD. The speech databases made within the OrienTel project were validated by SPEX, the Netherlands, to assess their compliance with the OrienTel format and content specifications.

Speech samples are stored as sequences of 8-bit 8 kHz A-law. Each prompted utterance is stored in a separate file. Each signal file is accompanied by an ASCII SAM label file which contains the relevant descriptive information.

Each speaker uttered the following items:

1 isolated single digit, 1 sequence of 10 isolated digits, 5 connected digits (1 prompt sheet number of 6 digits, 1 telephone number of 6-15 digits, 1 credit card number of 14-16 digits, 1 PIN code of 6 digits, 1 spontaneous phone number), 1 currency money amount, 2 natural numbers, 3 dates (1 spontaneous e.g. date or year of birth, 1 prompted date, 1 relative or general date expression), 2 time phrases (1 time of day spontaneous, 1 time phrase in word style), 3 spelled words (1 spontaneous e.g. own forename, 1 city name, 1 real word for coverage), 5 directory assistance utterances (1 spontaneous e.g. own forename, 1 city of childhood, 1 frequent city name, 1 frequent company name, 1 common forename and surname), 2 yes/no questions (1 predominantly ”yes” question, 1 predominantly ”no” question), 6 application keywords/keyphrases, 1 word spotting phrase using embedded application words, 4 phonetically rich words, 9 phonetically rich sentences.

The following age distribution has been obtained: 4 speakers are less than 16 years old, 179 speakers are between 16 and 30, 115 speakers are between 31 and 45, 29 speakers are between 46 and 60, and 5 are over 60 years old.
A pronunciation lexicon with a phonemic transcription in SAMPA is also included.

	ELRA members	Non-members
For research use	7,500 Euro	10,000 Euro
For commercial use	8,500 Euro	12,500 Euro

ELRA-S0160 Spanish Speecon Database

The Spanish Speecon database is divided into 2 sets:
- the first set comprises the recordings of 561 adult Spanish speakers (279 males, 282 females), recorded over 4 microphone channels in 4 recording environments (office, entertainment, car, public place).
- the second set comprises the recordings of 55 child Spanish speakers (27 boys, 28 girls), recorded over 4 microphone channels in 1 recording environment (children room).

The database has been collected by the Department of Signal Theory and Communications of the Universitat Politècnica de Catalunya (UPC) (Spain). The owner of the database is Siemens AG. This database is partitioned into 21 DVDs (first set) and 3 DVDs (second set).
The speech databases made within the Speecon project were validated by SPEX, the Netherlands, to assess their compliance with the Speecon format and content specifications. Each of the four speech channels is recorded at 16 kHz, 16 bit, uncompressed unsigned integers in Intel format (lo-hi byte order). To each signal file corresponds an ASCII SAM label file which contains the relevant descriptive information.
Each speaker uttered the following items:
- calibration data: 6 noise recordings, the “silence word” recording
- free spontaneous items (adults only): 5 minutes (session time) of free spontaneous, rich context items (an open number of spontaneous topics out of a set of 30 topics)
- 17 elicited spontaneous items (adults only): 3 dates, 2 times, 3 proper names, 2 city names, 1 letter sequence, 2 answers to questions, 3 telephone numbers, 1 language
- read speech: 30 phonetically rich sentences uttered by adults and 60 uttered by children, 5 phonetically rich words (adults only), 4 isolated digits, 1 isolated digit sequence, 4 connected digit sequences, 1 telephone number, 3 natural numbers, 1 money amount, 2 time phrases (T1: analogue, T2: digital), 3 dates (D1: analogue, D2: relative and general date, D3: digital), 3 letter sequences, 1 proper name

2 city or street names, 2 questions, 2 special keyboard characters, 1 Web address, 1 email address, 208 application specific words and phrases per session (adults), 74 toy commands and 48 general commands (children).
The following age distribution has been obtained:
- adults: 313 speakers are between 15 and 30, 176 speakers are between 31 and 45, 61 speakers are between 46 and 60, and 11 speakers are over 60.
- children: 19 speakers are between 8 and 10, 36 speakers are between 11 and 14.

A pronunciation lexicon with a phonemic transcription in SAMPA is also included.

	ELRA members	Non-members
For research use	50,000 Euro	60,000 Euro
For commercial use	67,000 Euro	75,000 Euro

ELRA-S0161 Russian Speecon Database

The Russian Speecon database is divsar, public place).
- the second set comprises the recordings of 50 child Russian speakers (31 boys, 19 girls), recorded over 4 microphone channels in 1 recording environment (children room).

The database has been collected by Auditech Ltd. (Russia). The owner of the database is Siemens AG. This database is partitioned into 21 DVDs (first set) and 3 DVDs (second set).
The speech databases made within the Speecon project were validated by SPEX, the Netherlands, to assess their compliance with the Speecon format and content specifications. Each of the four speech channels is recorded at 16 kHz, 16 bit, uncompressed unsigned integers in Intel format (lo-hi byte order). To each signal file corresponds an ASCII SAM label file which contains the relevant descriptive information.
Each speaker uttered the following items:
- calibration data: 6 noise recordings, the “silence word” recording
- free spontaneous items (adults only): 5 minutes (session time) of free spontaneous, rich context items (an open number of spontaneous topics out of a set of 30 topics)
- 17 elicited spontaneous items (adults only): 3 dates, 2 times, 3 proper names, 2 city names, 1 letter sequence, 2 answers to questions, 3 telephone numbers, 1 language
- read speech: 30 phonetically rich sentences uttered by adults and 60 uttered by children, 5 phonetically rich words (adults only), 4 isolated digits, 1 isolated digit sequence, 4 connected digit sequences, 1 telephone number, 3 natural numbers, 1 money amount, 2 time phrases (T1: analogue, T2: digital), 3 dates (D1: analogue, D2: relative and general date, D3: digital), 3 letter sequences, 1 proper name, 2 city or street names , 2 questions, 2 special keyboard characters, 1 Web address, 1 email address, 208 application specific words and phrases per session (adults), 74 toy commands and 48 general commands (children).

The following age distribution has been obtained:
- adults: 290 speakers are between 15 and 30, 187 speakers are between 31 and 45, 73 speakers are between 46 and 60.
- children: 28 speakers are between 8 and 10, 22 speakers are between 11 and 14.

A pronunciation lexicon with a phonemic transcription in SAMPA is also included.

	ELRA members	Non-members
For research use	50,000 Euro	60,000 Euro
For commercial use	67,000 Euro	75,000 Euro