Supports

Korterm, KAIST

Co-operating Organisations

ISO/TC37, Korterm, Infoterm, EAFTerm,
ISO/TC37/SC4

The Workshop Programme

14:30-14:45	Opening Address, Introduction and Summarization ----- *Christian Galinski, Key-Sun Choi*
14:45-15:00	General View of TC37/SC4 ----- *Laurent Romary*
15:00-15:20	General Methodology for TC37/SC4 ----- *Nancy Ide*
15:20-15:40	Terminology of Language Resources ----- *Klaus-Dirk Schmitz*
15:40-15:55	OpenNetTerminologyManager - a Web and Standards based OpenSource Terminology Management Tool ----- *Klemens Waldhör*
15:55-16:05	An XML Markup Language Framework for Lexical Databases Environments: the Dictionary Markup Language ----- *Mathieu Mangeot-Lerebours, Frédéric Andres*
16:05-16:10	Towards a generic architecture for Lexicon Management ----- *Cristina Vertan, Walther Von Hahn*
16:10-16:25	Management of Language Resources with Metadata ----- *Peter Wittenburg, Daan*

16:25-16:45	***Discussion***
16:45-17:00	***Coffee Break***
17:00-17:15	Towards Multimodal Content Representation ----- *Harry Bunt, Laurent Romary*
17:15-17:30	Where will the Standards for Intelligent Computer-Assisted Language Learning Come From? ----- *Lars Borin*
17:30-17:50	Standards for the Localization Industry ----- *Alan Melby*
17:50-18:05	Personal Names in Unrestricted Chinese Texts: Nature and Identification ----- *Benjamin K.TSOU, Lawrence Y.L.Cheung*
18:05-18:10	Changes in the Etymological Type of New Terminology in Japanese - The Decrease of Sino-Japanese and Increase of Alphabetical Terms- ----- *Takehiro Shioda*
18:10-18:15	A Corpus-based Approach to Term Bank Construction ----- *Bai Xiaojing, Hu Junfeng, Chen Yuzhong, Yu Shiwen*
18:15-18:35	***Discussion***

Workshop Organisers

- Laurent Romary, Laboratoire Loria, France
- Christian Galinski, Infoterm, Austria
- Nancy Ide, Vassar College, USA
- Key-Sun Choi, KAIST, Korterm, Korea

Workshop Programme Committee

- Gerhard Budin, University of Vienna, Austria
- Nicoletta Calzolari, CNRS, Pisa, Italy
- Key-Sun Choi, KORTERM, KAIST, Korea
- Yuzuru Fujiwara, National Center for Industrial Information, Tokyo, Japan
- Christian Galinski, Infoterm, Austria
- Koiti Hasida, Cyber Assist Research Center, Tokyo, Japan
- Gerhard Heyer, Leipzig University, Germany
- Isahara Hitoshi, CRL, Japan
- Junfeng Hu, Peking University, China
- Churen Huang, Academia Sinica, Taiwan
- Nancy Ide, Vassar College, USA
- Yeon-Bae Kim, Human Science Division, NHK, Japan
- Jong-Hyeok Lee, Postech, Korea
- Fang Qing, CNIS, China
- Laurent Romary, Laboratoire Loria, France
- Klaus-Dirk Schmitz, Fachhochschule Koeln, Germany
- Takehiro Sioda, NHK Broadcasting Culture Research Institute, Japan
- Virach Sornlertlamvanich, NECTEC, Thailand
- Tokunaga Takenobu, TIT, Japan
- Benjamin Tsou, City University of Hong Kong
- Sue-Ellen Wright, Kent State University, USA
- Shiwen Yu, Peking University, China
- Antonio Zampolli, CNRS, Pisa, Italy

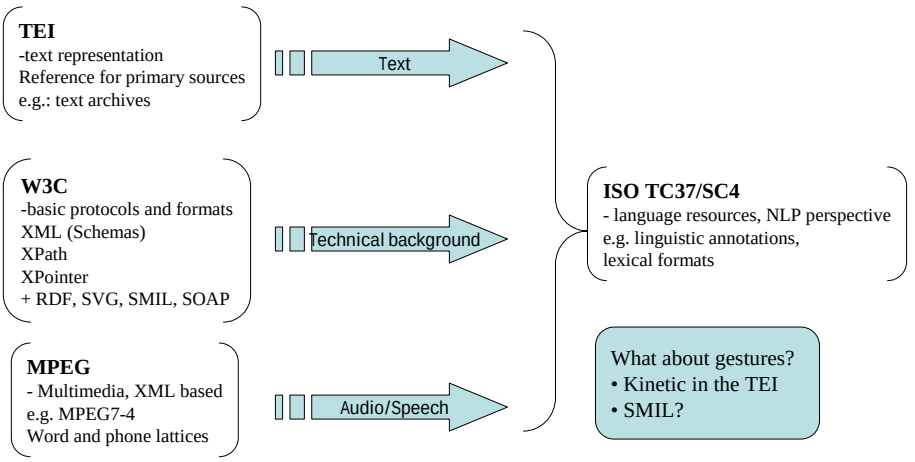
Table of Contents

Organization

- Secretary: Key-Sun Choi
- Provisional Chair: Laurent Romary
- International Advisory Committee
  - Permanent Chair: Prof. Antonio Zampolli

SC4 - environment

SC4 and other standardizing bodies



Possible sources for SC4

- Eagles, Mate
- Isle
  - Meta-data
  - Multilingual lexica
- OLIF
- OLAC
- MMA?

TC37/SC4 Work Items

- WI-1: Linguistic annotation framework
- WI-2: Linguistic resource documentation
- WI-3: Structural content representation scheme
- WI-4: Multimodal content representation sheme
- WI-5: Discourse level representation scheme
- WI-6: Multilingual text representation scheme

WI-1

- Linguistic annotation framework
  - Basic mechanisms and data structures for linguistic annotation and representation [data architecture]
    - Structural nodes and information units
    - Data category specification
    - Methods and principles for the design of an annotation scheme
    - Linking mechanisms
    - Feature Structures
  - Possible sources:
    - TMF, iso12620-revised, Mate (general methodology)
    - TEI (Linking mechanisms, feature structures)

WI-2

- Multimodal and multilingual information documentation
  - Description of a meta-data representation scheme to document linguistic information structures
    - General content description
    - Local content description
  - Possible sources:
    - Mile, OLAC
    - Data category specifications…

WI-3

- Structural content representation scheme
  - Definition of two annotation/representation schemes for morpho-syntax and syntax, to be used for annotation and interchange puposes
    - Meta-model for morpho-syntactic annotation
    - Meta-model for syntactic annotation (lexicalized grammar, elementary trees, dependancy structures)
    - Data category registry for morpho-syntactic annotation
    - Data category registry for syntactic annotation
  - Possible sources:
    - Eagles
    - TAGML
    - Working group with representatives from existing TreeBanks initiatives

WI-4

- Multimodal content representation sheme
  - Representation scheme for the integration of the semantic content of multimodal information (spoken, graphical and gestural)
    - Meta-modal for contant representation (Events, participants)
    - Data category registry for multimodal content
  - Possible sources:
    - SIGSEM working group on semantic content

WI-5

- Discourse level representation scheme
    - Meta-model for discourse and dialogue representation
    - Meta-model for discourse level annotation (e.g. reference annotation)
  - Possible sources:
    - Mate

WI-6

- Multilingual text representation scheme
  - Framework for representing bi- or multi-lingual textual information
    - Translation Memory
    - Alignment – Parallel Corpora
  - Possible sources:
    - TMX for translation memories
    - TEI based linking mechanism (or see WI-1) for Parallel texts

LREC Thematic session

- Special session on linguistic resource representation (chair K-S Choi)
  - Submitted papers - in concertation with LREC program committee
  - 30-45 minutes open discussion on main priorities for linguistic resource standardization

LREC Workshop

- Standardizing Linguistic Resources - Past activities & new prospects
  - Submitted papers
  - Round table + discussion on the definition of the work item, possible sources, etc.

Contacts

- DE: Alexander Geyken (Annotation schemes), Günter Neumann
- SP: Nuria Bel (POS/Syntax)
- NL: Harry Bunt (Semantics, SIGSEM)
- JP: Hashida Koichi

Terminology and Data Categories of Language Resources

Klaus-Dirk Schmitz
University of Applied Sciences, Cologne, Germany

Terminology Standards

Two meanings of "Terminology Standard"

- "Vocabularies" are terminology standards that contain subject-field-specific concepts and terms produced by terminology sub-committees on national, regional and international level
- Terminology-principles-and-methods standards produced by specific committees on national and international level (ISO/TC37, DIN-NAT, ...)

TC 37
Terminology and other language resources

Example 1:

Vocabularies

Principles and Methods

04/2002 Klaus-Dirk Schmitz 3

TC 37
Terminology and other language resources

- **ISO 1087-1**: Terminology - Vocabulary - Part 1
- **ISO 1087-2**: Terminology work - Vocabulary - Part 2: Computer applications

04/2002 Klaus-Dirk Schmitz 4

Contents

3 Vocabulary

3.1 Language and reality

3.1.1
object
anything perceivable or conceivable

NOTE Objects may be material (e.g. an engine, a sheet of paper, a diamond), immaterial (e.g. conversion ratio, a project plan) or imagined (e.g. a unicorn).

3.1.2
subject field
domain
field of special knowledge

NOTE The borderlines of a subject field are defined from a purpose-related point of view.

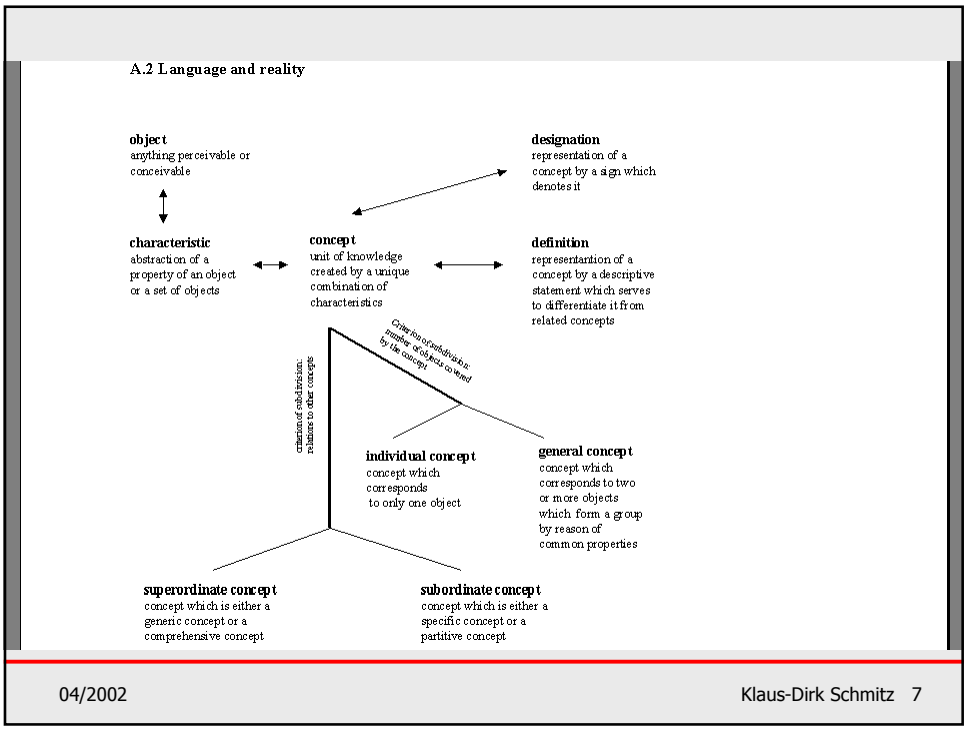
3.1.3
special language
language for special purposes
LSP
language used in a **subject field** (3.1.2) and characterized by the use of specific linguistic means of expression

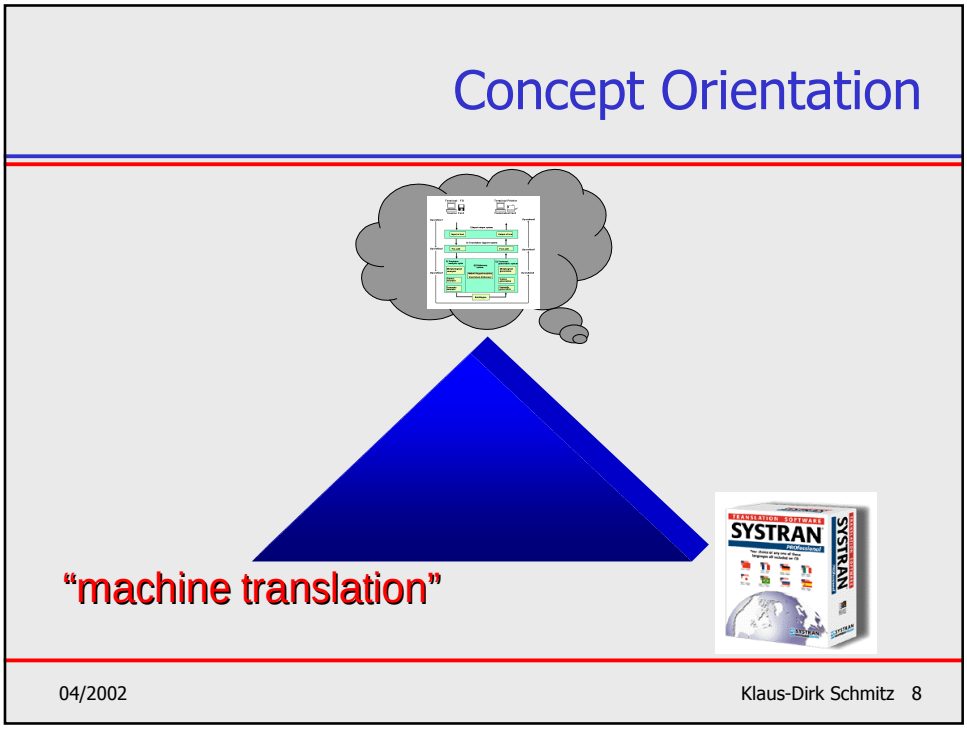
NOTE The specific linguistic means of expression always include subject-specific **terminology** (3.5.1) and phraseology and also may cover stylistic or syntactic features.

3.2 Concepts

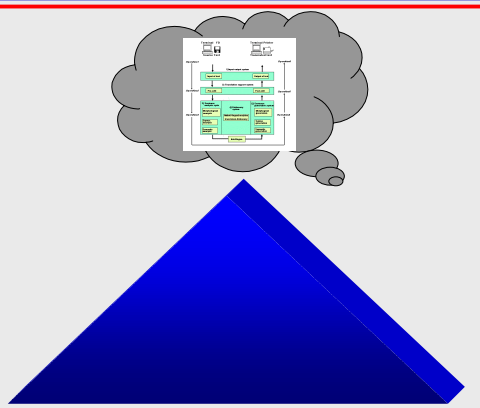
3.2.1
concept
unit of knowledge created by a unique combination of **characteristics** (3.2.4)

NOTE Concepts are not necessarily bound to particular languages. They are, however, influenced by the social or cultural background often leading to different categorizations.


A.2 Language and reality
object
anything perceivable or conceivable
designation
representation of a concept by a sign which denotes it
characteristic
abstraction of a property of an object or a set of objects
concept
unit of knowledge created by a unique combination of characteristics
definition
representation of a concept by a descriptive statement which serves to differentiate it from related concepts
Criterion of subdivision: number of objects covered by the concept
criterion of subdivision: relation to other concepts
individual concept
concept which corresponds to only one object
general concept
concept which corresponds to two or more objects which form a group by reason of common properties
superordinate concept
concept which is either a generic concept or a comprehensive concept
subordinate concept
concept which is either a specific concept or a partitive concept
04/2002
Klaus-Dirk Schmitz 7


Concept Orientation
"machine translation"
SYSTRAN
04/2002
Klaus-Dirk Schmitz 8

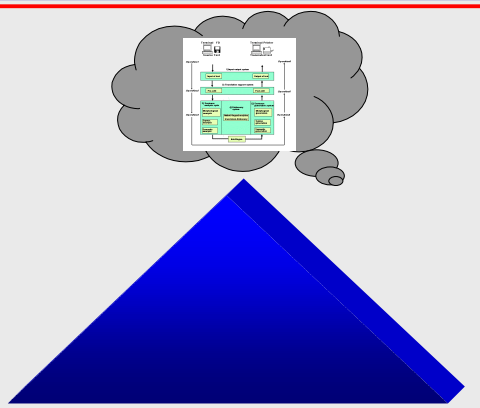
Concept Orientation



"machine translation" object

04/2002 Klaus-Dirk Schmitz 9

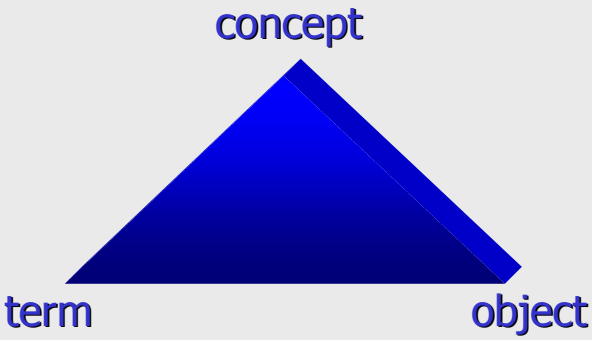
Concept Orientation



term object

04/2002 Klaus-Dirk Schmitz 10

Concept Orientation



Concept Orientation

object — Any part of the perceivable or conceivable world
Objects may be material (e.g. engine) or immaterial (e.g. magnetism)

concept — Unit of thought made up of characteristics that are derived by categorizing objects having a number of identical properties
Concepts are not bound to particular languages. They are, however, influenced by social or cultural background

term — Designation of a defined concept in a special language by a linguistic expression
A term may consist of one or more words

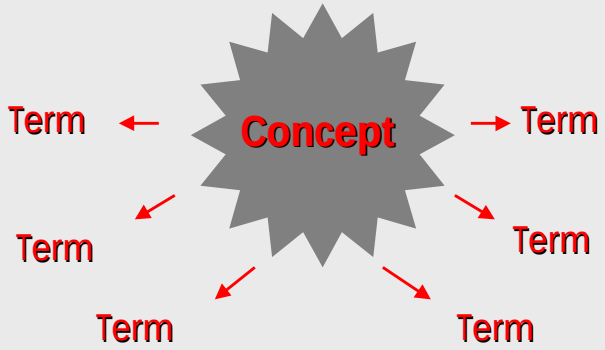
Concept Orientation

- All terminological information pertaining to *one* concept including all terms (designing this concept) in all languages and all descriptive and administrative data must be handled as *one* terminological unit.

04/2002 Klaus-Dirk Schmitz 13

Terminological Entry

Graphic adopted from Sue Ellen Wright

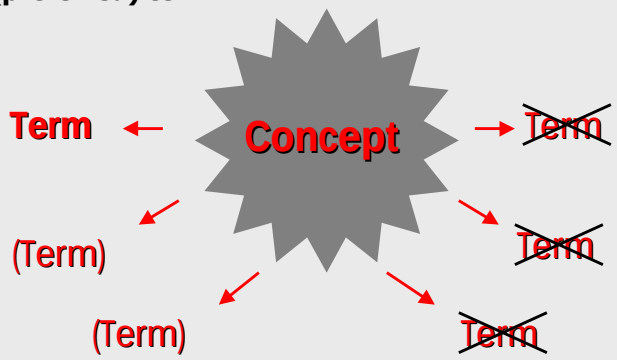


04/2002 Klaus-Dirk Schmitz 14

Terminological Entry

In standardized terminology:
only one (preferred) term !

Graphic adopted from
Sue Ellen Wright

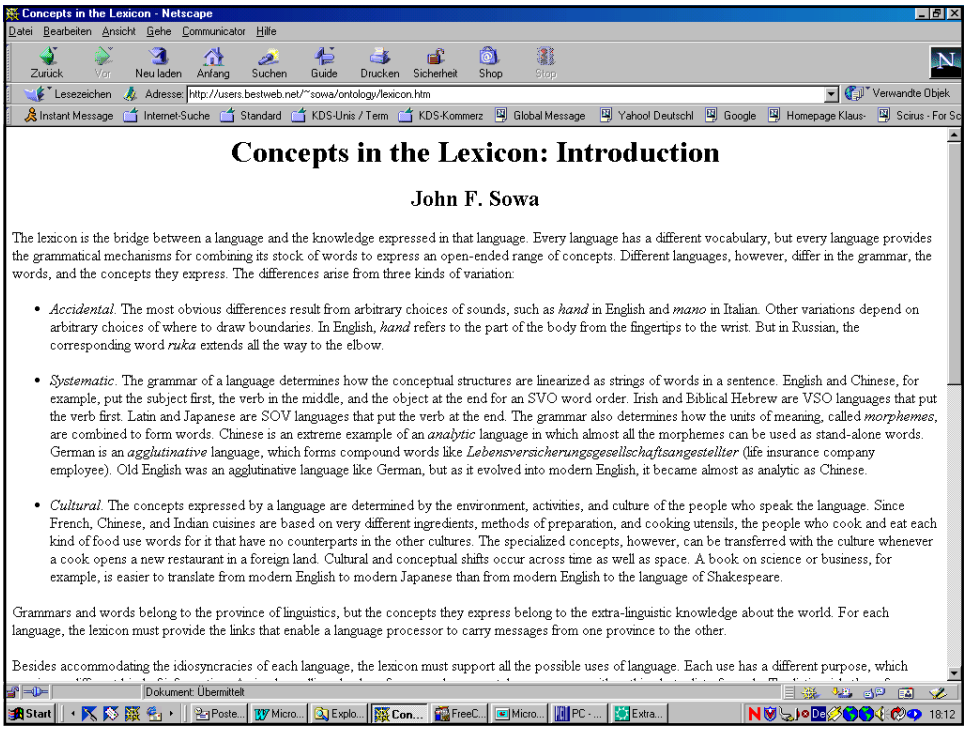


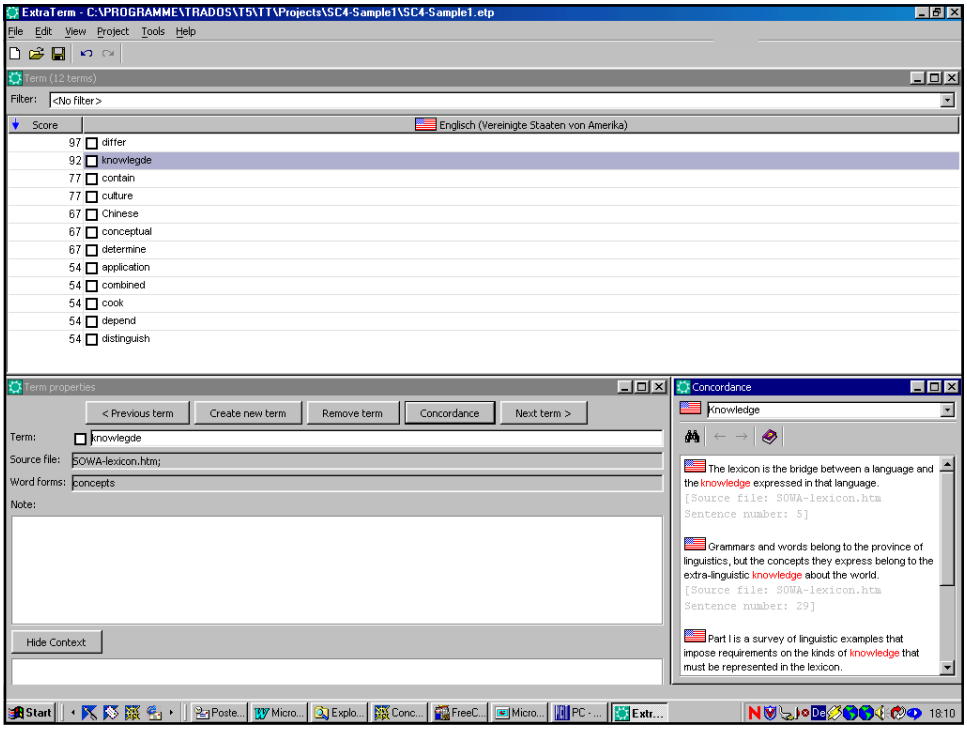
04/2002 Klaus-Dirk Schmitz 15

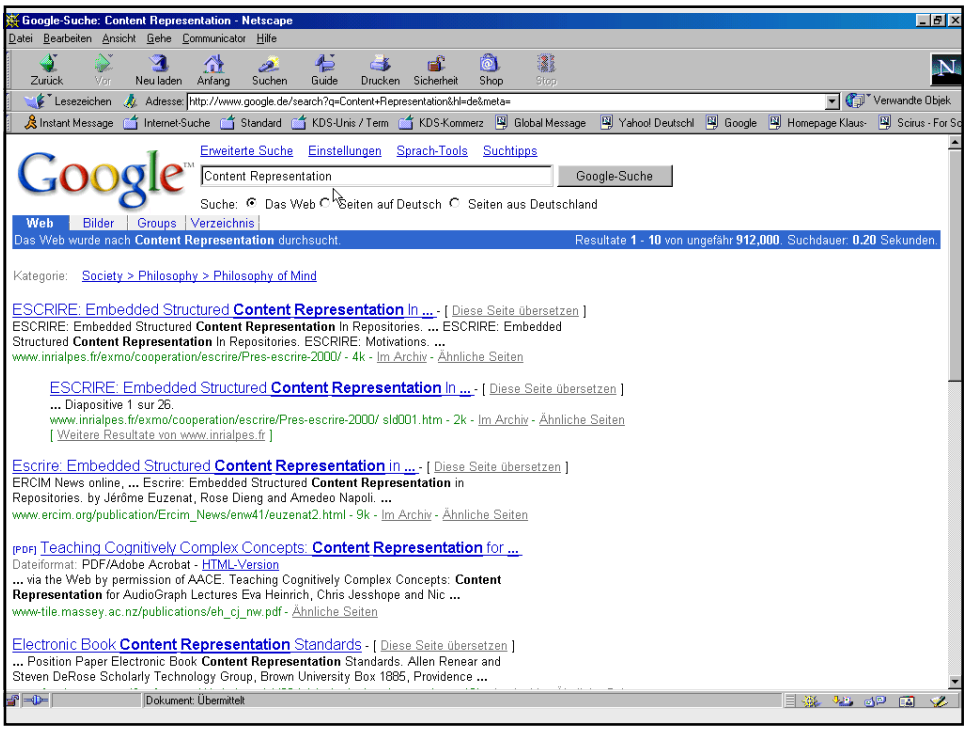
TC 37 / SC 4

- It is important to develop a standard for the terminology used in SC 4, i.e. the **terminology of language resources and language processing !**
- Preferably as the **first** work item in SC 4
- Data analysis material: LSP texts; websites
- Data analysis method: term extraction / term mining

04/2002 Klaus-Dirk Schmitz 16


Concepts in the Lexicon - Netscape
Datei Bearbeiten Ansicht Gehe Communicator Hilfe
Zurück Vor Neu laden Anfang Suchen Guide Drucken Sicherheit Shop Stop
Lesezeichen Adresse: http://users.bestweb.net/~sowa/ontology/lexicon.htm
Concepts in the Lexicon: Introduction
John F. Sowa
The lexicon is the bridge between a language and the knowledge expressed in that language. Every language has a different vocabulary, but every language provides the grammatical mechanisms for combining its stock of words to express an open-ended range of concepts. Different languages, however, differ in the grammar, the words, and the concepts they express. The differences arise from three kinds of variation:
Accidental. The most obvious differences result from arbitrary choices of sounds, such as hand in English and mano in Italian. Other variations depend on arbitrary choices of where to draw boundaries. In English, hand refers to the part of the body from the fingertips to the wrist. But in Russian, the corresponding word ruka extends all the way to the elbow.
Systematic. The grammar of a language determines how the conceptual structures are linearized as strings of words in a sentence. English and Chinese, for example, put the subject first, the verb in the middle, and the object at the end for an SVO word order. Irish and Biblical Hebrew are VSO languages that put the verb first. Latin and Japanese are SOV languages that put the verb at the end. The grammar also determines how the units of meaning, called morphemes, are combined to form words. Chinese is an extreme example of an analytic language in which almost all the morphemes can be used as stand-alone words. German is an agglutinative language, which forms compound words like Lebensversicherungsgesellschaftsangestellter (life insurance company employee). Old English was an agglutinative language like German, but as it evolved into modern English, it became almost as analytic as Chinese.
Cultural. The concepts expressed by a language are determined by the environment, activities, and culture of the people who speak the language. Since French, Chinese, and Indian cuisines are based on very different ingredients, methods of preparation, and cooking utensils, the people who cook and eat each kind of food use words for it that have no counterparts in the other cultures. The specialized concepts, however, can be transferred with the culture whenever a cook opens a new restaurant in a foreign land. Cultural and conceptual shifts occur across time as well as space. A book on science or business, for example, is easier to translate from modern English to modern Japanese than from modern English to the language of Shakespeare.
Grammars and words belong to the province of linguistics, but the concepts they express belong to the extra-linguistic knowledge about the world. For each language, the lexicon must provide the links that enable a language processor to carry messages from one province to the other.
Besides accommodating the idiosyncracies of each language, the lexicon must support all the possible uses of language. Each use has a different purpose, which
Dokument: Übermittelt
Start
18:12


ExtraTerm - C:\PROGRAMME\TRADOS\T5\TT\Projects\SC4-Sample1\SC4-Sample1.etp
File Edit View Project Tools Help
Term (12 terms)
Filter: <No filter>
Score
Englisch (Vereinigte Staaten von Amerika)
97 differ
92 knowlegde
77 contain
77 culture
67 Chinese
67 conceptual
67 determine
54 application
54 combined
54 cook
54 depend
54 distinguish
Term properties
< Previous term
Create new term
Remove term
Concordance
Next term >
Term: knowlegde
Source file: SOWA-lexicon.htm;
Word forms: concepts
Note:
Hide Context
Concordance
Knowledge
The lexicon is the bridge between a language and the knowledge expressed in that language.
[Source file: SOWA-lexicon.htm
Sentence number: 5]
Grammars and words belong to the province of linguistics, but the concepts they express belong to the extra-linguistic knowledge about the world.
[Source file: SOWA-lexicon.htm
Sentence number: 29]
Part I is a survey of linguistic examples that impose requirements on the kinds of knowledge that must be represented in the lexicon.
Start
18:10



TC 37 / SC 4

- It is important to use a web-based tool for managing terminology
- This tool must follow the basic principles of terminology management and the specific needs of ISO technical (sub-)committees standardizing the terminology of a specific domain
- In order to speed up the work between and in meetings

TC 37

Terminology and other language resources

Example 2:

Vocabularies

Principles and Methods

04/2002 Klaus-Dirk Schmitz 21

TC 37

Terminology and other language resources

- **ISO 12620:** Computer Applications in Terminology - Data categories
- **ISO 12200:** Computer Applications in Terminology - Machine-Readable Terminology Interchange Format (MARTIF) - Negotiated Interchange
- **ISO DIS 16642:** Computer applications in terminology – Meta model for representing terminological data collections / Terminology mark-up framework

04/2002 Klaus-Dirk Schmitz 22

ISO 12620 (Data Categories)

- Inventory of more than 200 data categories used in terminological data collections:
  - A.1 term
  - A.2 term-related information
  - A.3 equivalence
  - A.4 subject field
  - A.5 concept-related description
  - A.6 concept relation
  - A.7 conceptual structures
  - A.8 note
  - A.9 documentary language
  - A.10 administrative information
  - Annex B (informative): Bibliographical data categories

ISO 12620 (Data Categories)

A.2.2.1 part of speech

NONADMITTED TERM1**: grammatical category**

NONADMITTED TERM2**: word class**

DESCRIPTION: A category assigned to a word based on its grammatical and semantic properties.

PERMISSIBLE INSTANCES: Examples of parts of speech commonly documented in terminology databases can include:

a) noun
b) verb
c) adjective

- On the basis of a study and analysis of a great variety of practical applications; can be amended

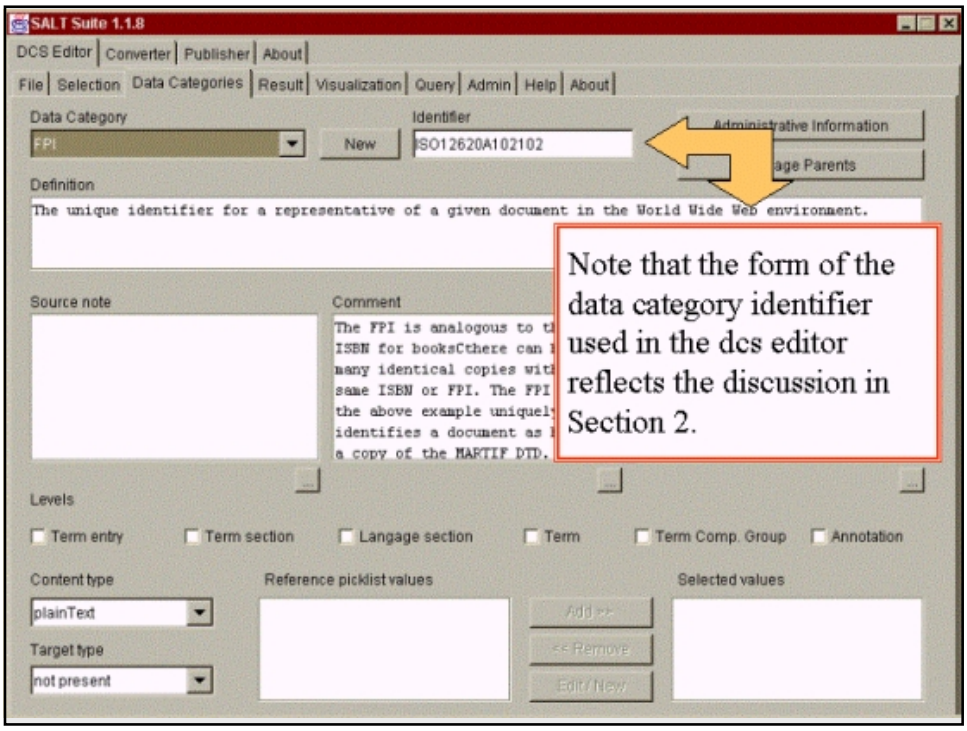
ISO 12620 new

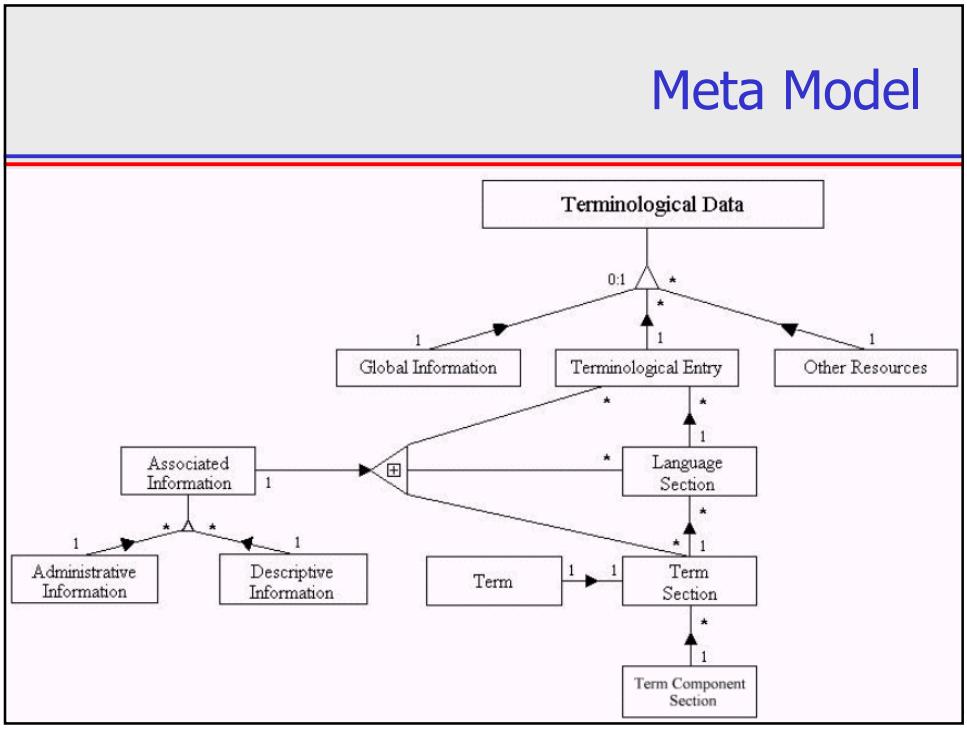
Metadata Registry

- contains terms that describe database fields
- for describing and comparing databases
- for human use
- "concept-oriented" but referring to objects (fields) that are IT representations of (real) objects/concepts
- ISO JTC1/TC32 provides a standard for metadata registries

ISO 12620 new

- Converting ISO 12620:1999 data category description into metadata registry format
- Using the DCS-Editor, developed within the framework of the SALT Project, for the description of the data categories
- Create the list of datCats and the description of datCats directly by the DCS-Editor as a normative annex of the new ISO 12620
- The body defines the (metadata) description format


SALT Suite 1.1.8
DCS Editor
Converter
Publisher
About
File
Selection
Data Categories
Result
Visualization
Query
Admin
Help
About
Data Category
FPI
New
Identifier
ISO12620A102102
Administrative Information
Definition
The unique identifier for a representative of a given document in the World Wide Web environment.
Source note
Comment
Note that the form of the data category identifier used in the dcs editor reflects the discussion in Section 2.
Levels
Term entry
Term section
Langage section
Term
Term Comp. Group
Annotation
Content type
plainText
Target type
not present
Reference picklist values
Selected values
Add >>
<< Remove
Edit / New


Meta Model
Terminological Data
0:1
Global Information
Terminological Entry
Other Resources
Associated Information
Language Section
Administrative Information
Descriptive Information
Term
Term Section
Term Component Section

ISO 12620 new

ISO 12620 Expanded Position	ISO 12620 Name	Target	DataType	Level Restrictions	Sub-set
A.02.02	grammar				DT
A.02.02.01	part of speech		plaintext	TS , TC	DT
A.02.02.02	grammatical gender		picklist 2	TS, TC	DT
A.02.02.03	grammatical number		picklist 3	TS, TC	DT
A.02.02.04	animacy		picklist 4	TS, TC	DT
A.02.02.05	noun class				DT
A.02.02.06	adjective class				DT
A.02.02.07.T	grammatical valency		plainText	TS	DT
A.02.02.08.T	inflection			TS, TC	DT

04/2002 Klaus-Dirk Schmitz 29

part of speech

Identifying and Definitional Attributes

Data Element ID:	ISO12620A020201 **Version No :** 1
Data Element Name:	part of speech
Type :	Data Element
Status :	Current 12-DEC-1999
Admitted Name:	
Non-admitted Name 1:	grammatical category
Non-admitted Name 2:	word class
Definition:	A category assigned to a word based on its grammatical and semantic properties.
Source-related Comment:	
Concept-related Comment:	
Example:	
Dictionary ID :	A.2.2.1

Relational and Representational Attributes

Datatype :	Plain Text (or user-defined picklist)
Representational Form :	??
Representation Layout :	??
Minimum Size :	
Maximum Size :	
Guide for Use :	In a given database, it is wise to configure this category as a user-defined picklist in order to avoid the proliferation of alternate forms, etc. For a global interchange format, however, it is important to specify this item as plainText because it is impossible to predict all the possible options that might occur in all possible language combinations
Validation Rules :	

Data Domain Details

Examples of parts of speech commonly documented in terminology databases can include:

Permissible value	Domain Meaning Definition Text	Example
noun	A word that refers to a person, place, thing, event, substance or quality.	'Doctor', 'tree', 'party', 'coal' and 'beauty' are all nouns.
verb	A word or phrase that describes an action, condition or experience.	The words 'run', 'keep' and 'feel' are all verbs.
adjective	A word that describes a noun or pronoun.	'Big', 'boring', 'purple', 'quick', 'obvious' and 'silvery' are all adjectives.
etc.		

TC 37 / SC 4

- It is important to develop a standard for data categories used in typical "SC 4 applications"
- May-be with different parts for different types of language resources (NLP lexica, texts, speech etc.)
- Data analysis material: language data collections
- Data analysis method/tool: (mod.) SALT's DCS-Editor

Thank you for your attention!

- MARTIF, ISO 12620 Data Categories, MSC, Meta-Model, TBX: www.ttt.org
- SALT Project: www.loria.fr/projets/SALT

klaus.schmitz@fh-koeln.de

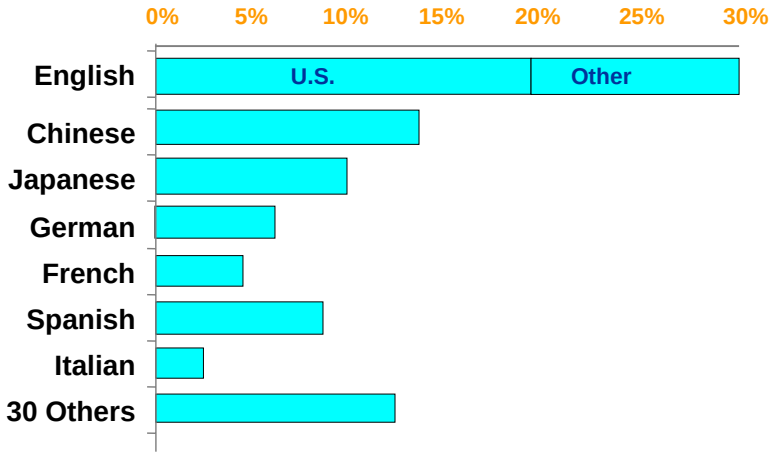
2002 May 28

Language Resource Representation and Management Standards for the Localization Industry

Alan K. Melby
<akm@byu.edu>

ttt.org -translation, theory & technology-

Tidbit: Internet Users by Native Language - 2005



Source: IDC Internet Commerce Market Model, V7.1 (c/o Rose Lockwood)

Overview of Presentation

A. Definition of localization as part of GIL
B. Brief history of LISA and OSCAR
C. Layers of Localization standards
D. XLIFF for text and source code
E. TMX for translation memory exchange
F. TBX and OLIF for terminology
G. Unresolved issue: segmentation

[A] GIL

- Globalization (G11N)
- Internationalization (I18N)
- Localization (L10N)

[A.1] Globalization

- Globalization (G11N)
  - Globalization is the business process of taking products and services into various new markets around the globe
  - A locale is the geographic region and language of a particular market
- Internationalization (I18N)
- Localization (L10N)

[A.2] Internationalization

- Globalization (G11N)
- Internationalization (I18N)
  - Internationalization is the engineering process of generalizing a product or service so that it can handle multiple languages and cultural conventions

Localization (L10N)

[A.3] Localization

- Globalization (G11N)
- Internationalization (I18N)
- Localization (L10N)
  - Localization is the cross-cultural communication process of preparing locale-specific versions of a product or service and consists of translation of textual material and adaptation of non-textual material.

[B] Brief History of LISA and OSCAR

- LISA (Localization Industry Standards Association) – see http://www.lisa.org/info/about.html
- OSCAR (Open Standards for Container/content Allowing Re-use)
- OSCAR is a LISA special interest group for language resource data standards

Localization-related Technologies

- Text Representation (Unicode and XML)
- Translation/Localization Container (TLC)
- Translation Tools (specialized)
  - Segmentation, alignment, encapsulation
  - Termbase setup or enrichment
  - Translation memory and machine translation
  - Terminology lookup
  - Missing segment and markup check
  - Term check (consistency, false friends, and variants)

[C] Layers of Localization Standards

- Unicode
- XML (including language/locale ids)
- XLIFF
- TMX
- TBX and OLIF

[D] XLIFF

- XLIFF is a format to store extracted text and carry the data from one step to another in the localization process
- see http://www.opentag.com/xliff.htm for more information

[E] TMX

- The purpose of TMX is to allow easier exchange of translation memory data between tools and/or translation vendors with little or no loss of critical data during the process.
- See http://www.lisa.org/tmx/ for more information

[F] TBX and OLIF

- TBX and OLIF allow the representation and exchange of terminological data, with a focus on human-oriented and NLP-oriented data, respectively
- See http://www.lisa.org/tbx/ for more on TBX
- See http://www.olif.net/ for more on OLIF

[G] Unresolved Issue

- Segmentation of words, sentences, and other linguistic units has not yet been standardized for the localization industry
- This means that word counts are not standard
- This also means that translation memory lookup may not detect some matches

OpenNetTerminologyManager- a Web and Standards based OpenSource Terminology Management Tool

Klemens Waldhör*

* Friedrichstr. 17, 90574 Roßtal, Germany, dr.klemens.waldhoer@waldhor.com

Abstract

OpenNetTerminologyManager is a privately started Open Source project which aims at developing a freely available pure web based concept terminology management system. It runs with any browser supporting JavaScript. The server side requires MySQL, Apache Web Server and Perl. The system is currently available through sourceforge.net at http://openwebterm.sourceforge.net. OpenNetTerminologyManager supports different terminological models. A version which is based on MARTIF has been implemented.

1. Introduction

Through the years the world has seen the attempt to establish several different terminology standards starting from MARTIF, Geneter to TBX, XLT (SALT), TMF (ISO 16642) and so on. The author himself was part of one of the older efforts which started 1990 where within the MULTILEX project a first try was made to create a standard exchange description for areas like mono- and multilingual dictionaries, machine translation etc. The basic idea there was to use SGML as the description language. This was followed up in projects like EAGLES, Otelo (OLIF) etc. In parallel other attempts have been made like Geneter. Sometimes one is really puzzled how creative the terminology community is in inventing new ideas and standards. Often it is really hard to follow what is going on. This is the one side of the coin. On the other side the industry uses "**quasi standards**" like the export format used in MultiTerm™ from Trados™. Several products of competitors like TermStar™, UniTerm™ and others provide import and export features from and into the MultiTerm™ format, simply because MultiTerm™ is the market leader in this area. Otherwise getting into this application field for new systems is nearly impossible as most customers either use MultiTerm™ or at least provide their data in this format.

Interestingly enough Open Source terminology software was never really part of the terminology game, in contrast to other areas like web servers where open source software like Apache is the dominating software (60 % of the world web server market). If one searches for "open source terminology management" in Yahoo and inspects the returned results in detail there are only two other relevant matches, the ForeignDesk and OpenGALEN match. In the last half year Lionbridge has made its software **ForeignDesk** available through open source. Another notable effort is **RosettaWerks** which deals implementing a set of tools for the localisation process.

But what is really missing is a terminology tool which is available on several operating systems (not just Windows™) and can be used through the web and itself is built on free available software. This is not the place to discuss the advantages of the open source model. A lot of discussion is going on this area, but I just want to add that one clearly has to distinguish the open source model from models which are offered by software suppliers where one can get the executables for free, but has no access to the source code. Several providers of terminology software supply down-graded or full versions of their tools – mainly viewers - e.g. UniLex™ from Acolada GmbH, but this does not bring any advantage to the user as he still relies on the provider to fix bugs etc. In addition it is hard to check if there are any hidden traps in the software. As professional terminology management contains company or customer information security aspects and the ability to check this will be an important aspect of choosing a system in the future. Based on this observations – and being also a fan of the open source community - I started developing a terminology management software which should fill this gap.

2. OpenNetTerminologyManager Terminology Model

The basic idea of the system architecture is the capability to support **different terminology models**. The user should have the option either to create his own model or to adapt an existing model by sub-classing it or adding his own fields. It should also be possible to keep track with on-going changes in the standardisation community. This has been realised in the system in the following way: attributes (elements) of the terminology model are not directly mapped to database tables, but this information is kept in a specific column where the structure can be freely defined. The actual mapping of these content of this column to attributes is defined in **model files**. Each database represents one model. The advantage of this approach is a) that it keeps the number of databases tables to a minimum, b) as a result the system is quite fast in searching and reading entries and c) adaptations of attributes can be made easily.

The basic OpenNetTerminologyManager approach is **concept oriented** as it used in most modern terminology systems. In this approach a concept corresponds to one meaning of a word. The language specific parts of a concept are called "**language terms**" or simply "**terms**". Each concept is tagged with an unique identifier, while each term related to the concept uses the concept identifier plus a language identifier and an internal term counter as identifier.

Example: The German term "**Birne**" (three meanings: Glühbirne, Frucht, Kopf = bulb, pear, nut) will be represented by creating three concepts (Figure 2):

a) one with the meaning of "Frucht = fruit" and

b) one with the meaning of "Glühbirne = bulb" and

c) one with the meaning of "Kopf = head".

The **kernel** of OpenNetTerminologyManager consists of several tables:

a) A **MONOTERM** table which holds all relevant information for a term including term attributes

b) A **MULTITERM** table which links entries in the MONOTERM table to a concept and also also stores concept related attributes.

c) A **DETAILS** table which contains links from attributes to terms and concepts. This table is only used to optimise the speed when searching with attributes.

d) A **LINK** table which establishes links between either concepts or term (e.g. in order to express a relations like "synonym").

Different terminology models are now mapped to the kernel model in a **model file**. This model file defines:

The **names** (e.g. "Gender") to be used for the **attributes** of the terminology model into an internal name. This association differentiates between concept related and term related attributes.

The **values** and **forms** to be associated with a such **names**. As an example associate the attribute "Gender" with three possible values ("male", "female", "neuter") and display them in the browser as a select box.

Table 1 shows a simple section for the MARTIF model. Models can further be differentiated into two classes: "**full models**" and "**sub-models**". A sub-model is defined as a subset of attributes of a full model. This is mainly necessary if for a given model (e.g. MARTIF) only specific attributes should be shown or if specific restrictions may apply for attribute values. The system contains some additional fixed attributes like the owner of the concept, read and write accesses etc.

3. OpenNetTerminologyManager Features

The following functions are currently supported:

- Constraints between attributes can be realised with JavaScript
- New models and sub-models can be created by the user (see Figure 1).
- Attributes can be defined by the user.
- Different types for attributes like option fields, text fields, select etc. are supported.
- Multiple databases; multi-user read/write support (locking at concept level). Different right combinations can be used. Databases are either private (with user and password protection) or public.
- Partial Unicode support. Unicode characters above Ascii 255 are stored as SGML entities in the database. This will be removed once MySQL supports directly UTF8 or a similar Unicode encoding scheme. Languages like Arabic, Chinese, Japanese etc. can be used through this approach. Once a Unicode implementation of MySQL is available this representation will be changed to an internal Unicode character set.

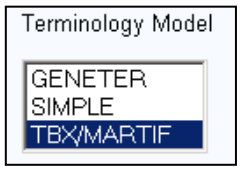


Figure 1: Models

Currently one terminology model based on MARTIF has been (partially) implemented. It normalises the XML definitions into the relational (table based) approach defined above. Others like Geneter are under way.

```
opwdetail40=Grammatical Gender
opwdetail41=Term Type
opwdetail44=Grammatical Number
…
 <tr>
   <td>
     <fieldset>
       <legend>Grammar</legend>
       <table>
         <tr>
           @tdopwdetail43=Part Of Speech?10<select!noun|verb|adjective|other
           @tdopwdetail40=Grammatical Gender?10<select!na|feminine|masculine|neuter|other
         </tr>
         <tr>
           @tdopwdetail41=Term Type?10<select!...|variant
           @tdopwdetail46=Valency?10<input
         </tr>
         <tr>
           @tdopwdetail44=Grammatical Number?10<select!na|dual|mass|other|plural|singular
           @tdopwdetail45=Animacy?10<select!animate|inanimate|other
         </tr>
       </table>
     </fieldset>
   </td>
 </tr>
```

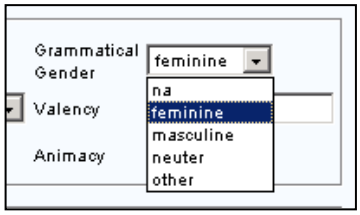


Table 1: OpenNetTerminology Manager GUI description

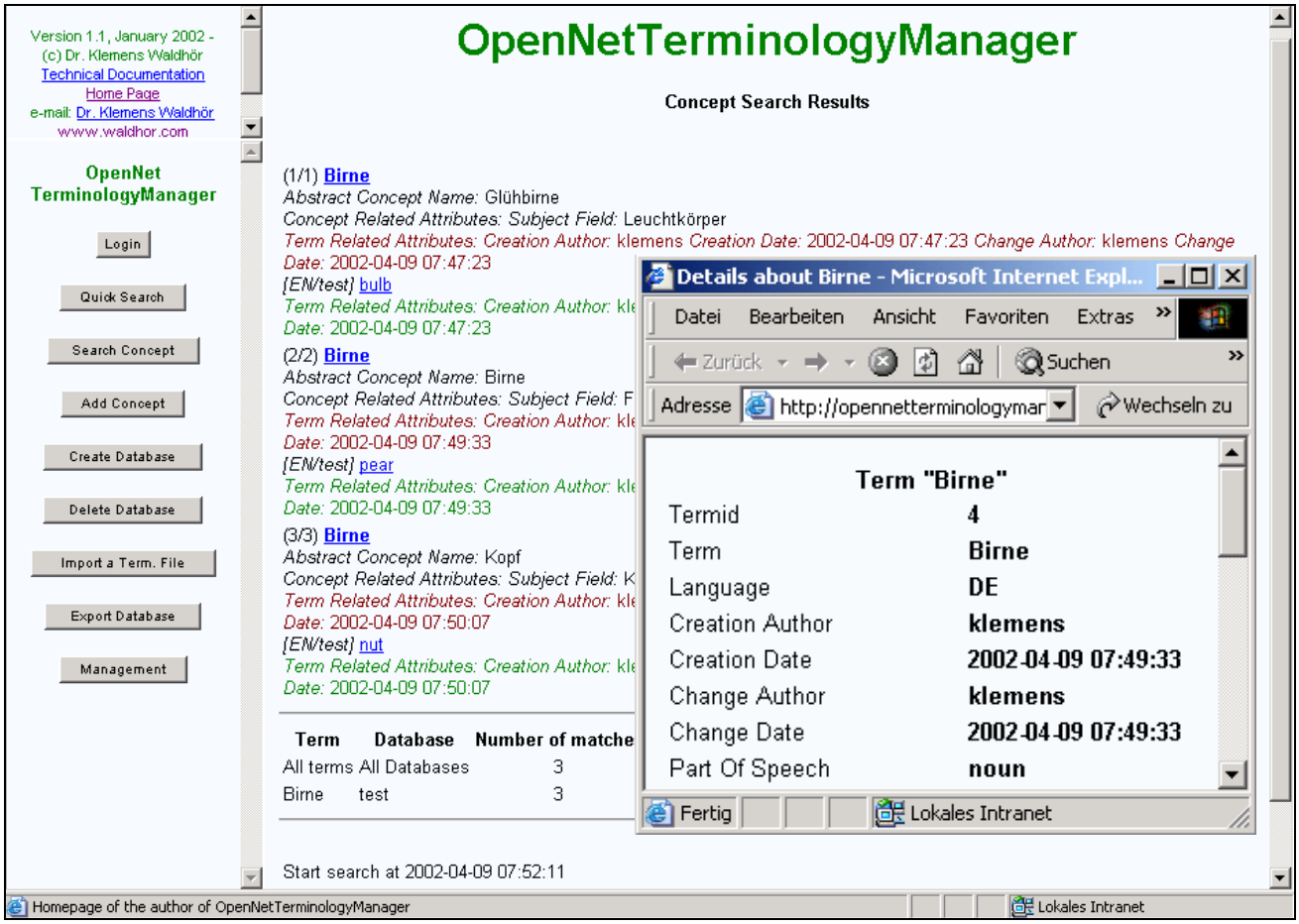
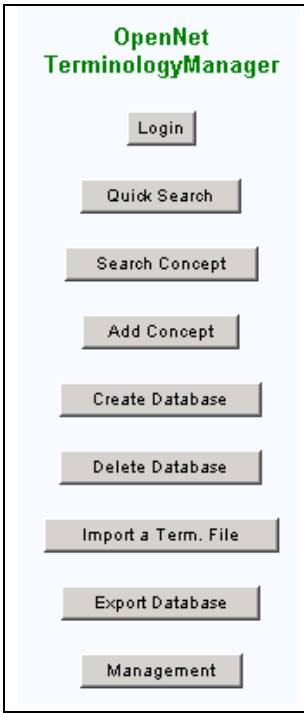


Figure 2: OpenNetTerminology Manager User Interface



Login allows the user to define default values like his preferred databases, languages, attributes to be displayed for a model, how search results should be displayed etc.

Quick Search offers a simplified search mode which simply looks up the database for a specified term independently of the language.

Search Concept supports more detailed search options like languages to be searched and displayed etc. Through this item the user also can edit concepts in the database.

Add Concept adds new concepts to the database.

The user can create new databases using **Create Database** where he also specifies the model and access rights for the database.

Databases can be deleted using **Delete Database**.

Import Term File allows existing terminology files to be imported in various formats supporting double detection during import,

while **Export Database** exports databases into various formats.

Management is mainly intended to give an overview of current system settings and databases. It also supports recreating the database structure.

Figure 3: OpenNetTerminologyManager Commands

4. The Basic User Interface of OpenNetTerminologyManager

Figure 2 shows the basic web based user interface. It consists of a main window where the results of queries etc. are shown and a navigation window (left). Optionally additional concept or term related information can be displayed in a separate browser window. Figure 3 describes the basic functions of the navigation window.

Concepts can be edited by first searching them with the **Search Concept** function and using the "Edit Mode" (not choosing "Dictionary View" option). See figure 4. Results are then displayed in a tabular like format (figure 5). Clicking on "Edit" will then display the full entry (figure 6) in an editable format. Results can also be displayed in a "Dictionary View" mode (figure 7). In this mode concepts found with the same name for a given language may optionally be collapsed into one output entry. This displays the entry in a similar way as they are show in printed dictionaries. Depending on the user search result display settings attributes will be displayed either directly in the main window as part of the entry or the term name is realized as a hyperlink and when clicking on it is displayed later in a separate browser window (figure 2). In addition the user can configure for each database which attributes should be shown. The query itself supports various search options like full text search, regular expressions, the LIKE operator etc.

5. Software requirements

OpenNetTerminologyManager requires the following software components: Perl > 5.0 (with some additional modules installed), Apache Web server or a compatible server, MySQL and a JavaScript enabled Web Browser. Tests have been done with Internet Explorer 5.0, 6.0®, Netscape® and Opera® . The system has been tested both on Windows (NT® and 2000®) and LINUX.

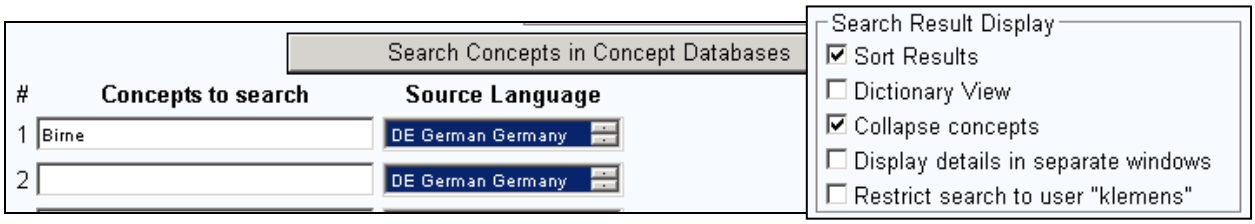


Figure 4: Searching concepts

No	ID / Concept	Source Language DE	Translation Term	Language	Database	Operation
1/1	316071404	Birne	nut	EN	meine	Edit Delete Copy Concept
2/2	284021368 Glühbirne	Birne	bulb	EN	meine	Edit Delete Copy Concept
3/3	312237593	Birne	pear	EN	meine	Edit Delete Copy Concept

Figure 5: Searching result display

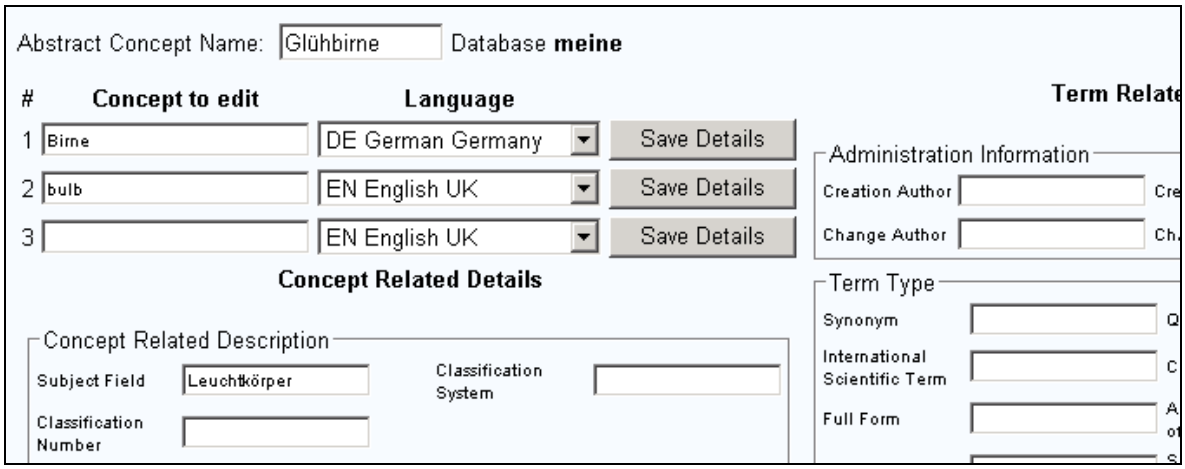


Figure 6: Editing concepts

Figure 7: Dictionary View display with no attributes displayed searching for "Planet%"

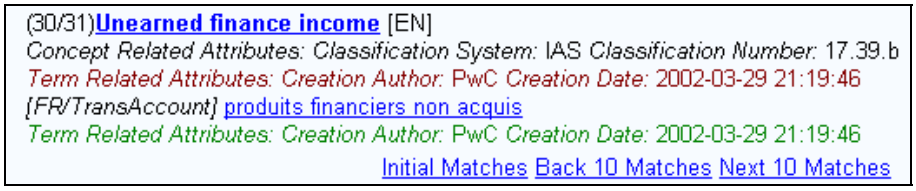


Figure 8: Result of a TransAccount terminology database full text query (searching for the term "finance") with attributes displayed.

6. Application Scenario

The TransAccount project (MLIS 5016) deals with the need for a multilingual translation system allowing the translation and interpretation between the annual accounts of a member state of the European Union (France) and IAS (International Accounting Standards) statements. Within this project the XBRL (eXtensible Business Reporting Language) IASCF taxonomy has been translated from English to French by one of the partners. The resulting 2000 concepts have been imported into a TransAccount terminology database. In addition about 2000 other general financial terms have been converted from a Geneter based format which have been produced by another partner at the start of the project. An example of the results of a query is shown in figure 7.

7. Next Steps

An important feature which is currently under development is an advanced link concept. This link concept will not only support links in the way as TBX defines them but will allow to create complex typed links between concepts and terms and databases. This will allow the user to search the databases not only as a simple term-lookup tool but to browse through it in a kind of semantic net and to find related concepts.

A concept is also developed which supports "similarity queries". It is intended to introduce a "stemming based index" by applying the Porter stemming algorithm to terms for some languages automatically (Porter, 1980). Other developments concern additional import / export formats and simplified form handling for attributes.

As there are several opens source project on mapping xml to relation databases on the way (e.g. XML-DBMS) I am currently also looking into replacing the internal structure of the database by a full xml database approach. This will heavily depend on the access speed compared to the current implementation.

8. References

Acolada. http://www.acolada.de

ForeignDesk. http://sourceforge.net/projects/foreigndesk/

OLIF. http://www.olig.net

OpenGALEN. http://www.opengalen.org/

OpenNetTerminologyManager. http://openwebterm.sourceforge.net

Porter, M.F., 1980. An algorithm for suffix stripping, *Program, 14 no. 3, pp 130-137, July 1980*

Rosettawerks. http://rosettawerks.sourceforge.net/Default.php

Sourceforge. http://sourceforge.net/

Star AG. http://www.star-ag.ch/eng/software.html

Trados. http://www.trados.com

TransAccount: http://www.transaccount.org

XML-DBMS. http://www.rpbourret.com/xmldbms/index.htm

Waldhör, K., Tesniere, B., 2002. Multilingual Terminology Database, *MLIS 5016 TransAccount Report*.

XBRL. http://www.xbrl.org

9. Acknowledgements

Thanks has to be given at this place to SourceForge which provides an excellent – and free – way to make open source projects available through the web.

An XML Markup Language Framework for Lexical Databases Environments: the Dictionary Markup Language.

Mathieu MANGEOT-LEREBOURS **Frédéric ANDRÈS**

Software Research Division, NII
Hitotsubashi, 2-1-2 Chiyoda-ku
101-8430 Tokyo, Japan
mangeot@nii.ac.jp

Introduction

Lexical data resources are growing rapidely thanks to the Internet. Unfortunately, despite numerous existing standards like TEI, MARTIF, GENELEX, EAGLES/PAROLE, etc. each resource has its own format and own structure. Furthermore, the existing lexical data is generally developed for a specific purpose and can't be reused easily in other applications.

In this paper, we intend to define a complete framework for developing multilingual lexical database for multipurpose. The framework is generic enough in order to accept a wide range of dictionary structures and proposes for manipulating heterogeneous dictionaries a set of common pointers into these structures.

We will first present the organisation of Dictionary Markup Language (DML) framework.

Then we will describe more precisely the DML language based on XML schemata.

Next, we explain how to describe dictionary macro and microstructures with the DML.

Lastly, we will explain our concept of common pointers defined in a Common Dictionary Markup (CDM) set.

1. Presentation of the DML Framework

The DML Framework described first by Mangeot-Lerebours (2001) is a complete framework for the consultation of heterogeneous dictionaries, cooperative construction of new dictionaries and communication with other lexical databases or lexical data client and supplier applications. The framework is completely generic in order to manage heterogeneous dictionaries with their own proper structures.

The consultation of heterogeneous dictionaries is possible as soon as they are encoded in XML, consultation of other resources via remote servers through API, possibility of adding pre-consultation help modules such as spell checking and morphological analysis before consultation or post-consultation modules like syntethisers, conjugation of verbs, learning drills, etc. Possibility of automatic consultation of the database via client API.

The construction of new dictionaries can be done by a community of contributors and validated by a group of head lexicographers specialists.

The management of user profiles, preferences and weights for consultation, annotation and edition of lexical data with inheritance and sharing possibilities among groups of users is also handled by the framework.

The `<database>` element describes a lexical database and lists the dictionaries that are stored in it.

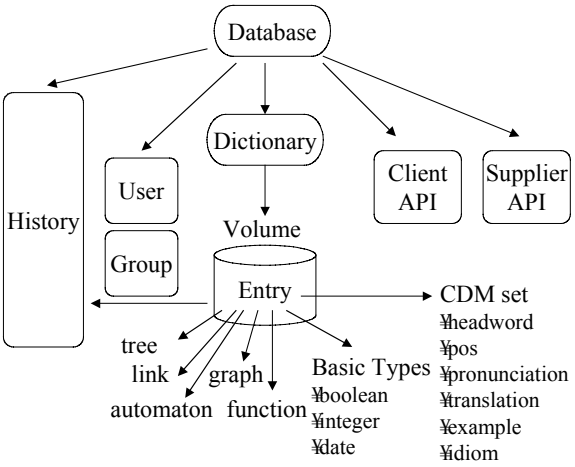


Figure 1. Logical Organisation of a Lexical Database

The `<dictionary>` element describes the metadata linked to ther dictionary. It links all the volumes of the dictionary.

The `<volume>` element describes a dictionary part. The content is principally a list of dictionary entries. For example, a bilingual bidirectional French-English dictionary will be described by only one `<dictionary>` element. The French->English entries will be in one `<volume>` element and the English->French entries in another `<volume>` element

2. The DML Language

2.1. The DML Namespace

To describe the structure of all the documents, elements, attributes and XML types, we use an XML namespace [XML Namespaces]. Our namespace is called DML for Dictionary Markup Language. The

namespace URI points to an XML schema [XML Schemas] describing the contents of the namespace. It is available online[1] to allow users to edit and validate their files online with an XML schema validator.

```
<MyElement
xmlns:dml="http://www-clips.imag.fr/
geta/services/dml">
  ...
  <dml:MyDescendant/>
  ...
</myElement>
```

Figure 2: Usage Example of the DML Namespace

2.2. DML Common Types and Attributes

For some information, we define type and attributes common to all DML elements. It allows to standardize the data. The XML schemata have originally simple predefined types. We selected and reused some in our definitions.

2.2.1. Dates and Time

Dates are represented by the `date` DML attribute of the XML schema type dateType taken from the extended format of the ISO 8601 standard.

2.2.2. Response Delay

The `delay` DML attribute of an element indicate the response delay when a request has been launched on this element.
This delay is a duration of the XML schema `durationType` type. For example, 5 seconds and 10 cents will be indicated : "5.10S".

2.2.3. Unique ID

The `id` DML attribute of an element is a unique ID in all the lexical database. It allows to create links between elements. It redefines the XML schema ID simple type.

2.2.4. Modifications History

The modifications history of an element has a unique ID. The element links to its history thanks to the DML attribute `history` that gives the value of the history ID. The type redefines the XML schema ID simple type.

2.2.5. Languages Notation

To note the various languages, we use the ISO-639-2/T (T for Terminology) [ISO98] standard that defines a 3 letter code for each language (French->fra; English->eng, Malay->msa, etc.). It is far more complete that the two letters code standard ISO-639-1. We also add our proper codes like "unl" for the UNL language. This codes list represents the lang DML type. The `lang` DML attribute is from this type.

[1] http://www-clips.imag.fr/geta/services/dml/

2.2.6. Documents Encoding

To note the encodings of the various documents in the database, we define the `encodingType`. DML type. The values are those described by the IANA (Internet Assigned Number Authority) for the encodings. These are also the values used for MIME types (Multipurpose Internet Mail Extension). Among the most used, we find ASCII on 7 bits, ISO-8859-1 on 8 bits for latin languages, Shift-Jis on 8 or 16 bits for the Japanese, UTF-8 on 8 bits for UNICODE characters, etc.

2.2.7. Status of an Element

The `status` DML attribute is used to indicate its status. The values can be among others `auto` if the element has been obtained automatically, `rough` if the element has not been revised and `revised` if so, etc.

3 DML Architecture

3.1. Macrostructure Definitions

To describe the macrostructure of our dictionaries as well as our lexical database, we use XML elements. We principally based our definitions on the LEXARD language defined by Serasset (1994) and added some information

3.1.1. Description of a Lexical Database

To describe a lexical database, we use the `<database>` element formally described inthe DML schema.
The modifications of the `<database>` element and its descendants are stored in a document linked with the `history-ref` atttribute.
We add to LEXARD the possibility to define various users and groups in the database. At the beginning three groups are predefined : `universe` contains all the users of the database, `administrators` contains the administrators of the database and `lexicologists` contains the users in charge of the control of the data.
The information relative to each user are stored in another element referenced by the `<user-ref>` element.
All the dictionaries of the database are referenced by pointers on XML documents that describe them. The pointers are the `href` attributes of the `<dict-ref>` elements grouped in the `<dictionaries>` element.

3.1.2. Description of a Dictionary

To describe a dictionary, we use the `<dictionary>` element. The modifications information is stored in a document pointed by the `history-ref` attribute.
We indicate meta-information on the resources.
The elements `<category>`, `<type>` and `<links>` describe the dictionary macrostructure.

The <category> element indicates the dictionary type (monolingual, bilingual, multilingual, interlingual). The <type> element indicates if the dictionaries are unidirectional, bidirectional or pivot based.
The <links> element indicates the links between the volumes of the dictionary. For example, if a dictionary is pivot based with 3 languages English, French and Malay, it contains 4 volumes Interlingual, English, French and Malay linked as follows:

```
<links>
  <link from="English"
to="Interlingual"/>
  <link from="French"
to="Interlingual"/>
  <link from="Malay"
to="Interlingual"/>
</links>
```

The dictionary volumes are referenced by their unque name. The <volumes> element gathers all the reference to the volumes files noted with the <volume-ref> element.
The source and target languages are indicated with the 3 letter code DML lang type.
The <content> element describes the content of the dictionary. The <domain> element indicates the domain covered by the dictionary (general, medecine, computer, etc.)
We indicate also the size of the dictionary in bytes by <bytes>, and the headword number by <hw-number>.
For the version management, we indicate the version number (<version>), the creation-date of the dictionary (<creation-date>) and the date of the integration of the dictionary into the database (<installation-date>).
For the non-DML resources, we need to indicate the file format (<format>) and the encoding (<encoding>). The encoding values are determined by the DML type encodingType.
We also indicate meta-information on the dictionary like the resource supplier (<source>), the owner (<owner>), the responsible at the database level (<responsible>), the rights attached to the dictionary (<legal>) and miscellaneous comments (<comments>).
The CDM (see chapter 4) elements list (<cdm-elements>) is stored with for each element, its real name in the resource and the maximal response delay. The (<corpus>) element is special, it allows to indicate that we search a string anywhere in the dictionary.

3.1.3. Description of a Volume

The <volume> elements gathers dictionary entries with the same source language. The modificaitons history is referenced with the history-ref attribute.

3.2. Microstructure Definitions

To represent dictionary microstructures, we propose to redefine in XML the structures defined with LINGARD (see serasset (1994).

3.2.1. Trees

To represent a dependance tree associated to the sentence "Le chat mange une souris.", for example, we can use a “decorated node” <dn> with attributes corresponding to the grammatical variables.

```
<dn ul="manger" time="present"
aspect="imperfectif">
  <dn ul="chat" determ="defini"
gnr="masc" pos="-1"/>
  <dn ul="souris" determ="indefini"
gnr="fem" pos="+1"/>
</dn>
```

3.2.2. Links

The definition of a link is done with the xlink standard [XLink 1.0]. We alslo add our attributes:

- The attribute type="bidirectionnal" or type="oriented" indicates if the link is bilingual or not;
- The attribute id is of the DML id type. It allows to attribute a unique id for each link;
- The content text of the element allows to tag the links.

Here is a link example:

```
<link type="oriented" id="l001"
href="example.xml#xpointer(//node[xl
:label='n002'])"/>
```

The reference to the external element is done with the href attribute. The reference is noted as a URI. If the object does not have a unique id (id), the link is described with the [XPointer] standard. Otherwise, it is pointed as follows:

```
<link type="oriented" id="l001"
href="example.xml#n002"/>
```

3.2.3. Graphs and Automatons

The xlink standard [XLink 1.0] is used to describe arcs. The arcs type is oriented type="oriented" or bijective type="bijective". The source and the target of the arc are noted with the node identifiers from="n001" and to="n002".
The definition of an automaton follows the definition of a graph. The starting node is noted with the xl:title="starting-node" attribute. The ending nodes are noted with the xl:title="ending-node" attribute.

3.2.4. Functions

The following example represents the lexical function [lambda]x1 ($CausOper_1$x0x1). The results of its application to the French lexie DÉSESPOIR are the following: pousser, réduire quelqu'un au désespoir,

jeter quelqu'un dans le désespoir, frapper quelqu'un de désespoir. The function is noted in XML as follows:

```
<function name="CausOper1">
  <arguments>
    <first value="desespoir"/>
  </arguments>
  <valgroup>
    <value>pousser</value>
    <value>réduire [qqun au
désespoir]</value>
    <value>jeter [qqun dans le
désespoir]</value>
    <value>frapper [qqun de
désespoir]</value>
  </valgroup>
</function>
```

3.2.5. Feature Structures

If the features are typed, the type is noted with an attribute. If the feature has several values, the element is duplicated.

```
<feature1
type="type1">valeur1</feature1>
<feature1
type="type2">valeur2</feature1>
```

3.2.6. Sets and Disjonction

Sets and disjunctions are defined directly at the XML schema level with the two elements <xsd:choice> and <xsd:sequence>

3.2.7. Basic Types

The basic type of an XML document is the character string. Thanks to XML schemata, we can use many other basic types like boolean, entity, decimal, float,etc.

4. The Common Dictionary Markup Subset

We defined a subset of DML element and attributes that are used to identify which part of the different structures represent the same lexical information. This subset is called Common Dictionary Markup (CDM).

4.1. Definition of the Subset

The DML framework may be used to encode many different dictionary structures. Indeed, two dictionary structures can be radically different. So, in order to handle such heterogeneous structures with the same tools, we need a common formalism. Standards like TEI [Ide95], MARTIF [Melby94], [ISO99]; GENELEX/EAGLES [GENELEX93] and [GENETER] aim to be universal but very few resources implement them.

We made a more pragmatic work with identifying the information in the existing resources as well as their meaning and naming them ina unique way in the DML namespace

This hierarchized subset is called Common Dictionary Markup and comes principally from the detailed examination of the FeM, DEC, OHD, OUPES, NODE, EDict, ELRA-MÉMODATA dictionaries and the 12th chapter of the TEI about dictionaries. It contains the most frequent elements found in these resources like the headword, the pronunciation, the part-of-speech, the examples, the idioms, etc. These elements have always the same semantics. For example, <dml:entry> always refer to a dictionary entry and <dml:headword> to the headword.

For some elements with closed lists of values, we define a list representing the intersection of the values and conversion rules for each resource. An example is the list of parts-of-speech for each language.

This set is in constant evolution. If the same kind of information is found in several dictionaries then a new element representing this piece of information is added to the CDM set. It allows tools to have access to common information in heterogeneous dictionaries by way of pointers into the structures of the dictionaries. The table 1 lists a first version of the CDM subset.

<CDM tag>	**(TEI equivalent)**
<entry>	(entry)
<headword hn="">	(hom)(orth)
<headword-var>	(oVar)
<pronunciation>	(pron)
<etymology>	(etym)
<syntactic-cat>	(sense level="1")
<pos>	(pos)(subc)
<lexie>	(sense level="2")
<indicator>	(usg)
<label>	(lbl)
<definition>	(def)
<example>	(eg)
<translation>	(trans)(tr)
<collocate>	(colloc)
<link href="">	(xr)
<note>	(note)

Table 1: CDM Elements Subset

4.2. CDM Correspondance Examples

When a resource is recuperated, a correspondance table is established between the original element names and CDM elements. The table 2 has been used for the FeM, OHD and NODE dictionaries.

CDM	FeM	OHD	NODE
<entry>	<fem-entry>	<se>	<se>
<headword>	<entry>	<hw>	<hw>

<pronunciation>	<french_pron>	<pr><ph>	<pr><ph>
<etymology>			<etym>
<syntactic-sense>		<sense n=1>	<s1>
<pos>	<french_cat>	<pos>	<ps>
<lexie>		<sense n=2>	<s2>
<indicator>	<gloss>	<id>	
<label>	<label>		<la>
<example>	<french_ sentence>	<ex>	<ex>
<definition>			<df>
<translation>	<english_equ> <malay_equ>		<tr>
<collocate>		<co>	
<link>	<cross_ref _entry>	<xr>	<xg> <vg>
<note>		<ann>	

Table 2: Equivalents of the CDM elements in the FeM, OHD and NODE

Conclusion

This framework has been extensively used for the Papillon project (see Serasset & Mangeot-Lerebours (2001)) of mutualized construction and consultation of a pivot multilingual lexical database. This experiments allowed us to correct and adapt some parts of the DML.
Nevertheless, the framework need to be opened to the public in order to receive feedback and comments. We plan to open a web site dedicated to the DML soon.

References

GENELEX (1993) Projet Eureka Genelex, modèle sémantique. Rapport Technique, Projet Eureka, Genelex, mars 1994, 185 p.

Nancy Ide & Jean Veronis (1995) Text Encoding Initiative, background and context. Kluwer Academic Publishers, 242 p.

ISO (1998) ISO 639-1 & 2 Code for the representation of names of languages Part 1 & 2 Alpha-3 code. Geneva, Part 1: 17 p., Part 2: 90 p.

ISO (1999) ISO DIS 12200 (MARTIF) Computer applications in terminology - Machine-readable terminology interchange format - Negotiated interchange.ISO TC 37/SC 3/WG I, Geneva, 118 p.

Mathieu Mangeot-Lerebours (2001) *Environnements centralisés et distribués pour lexicographes et lexicologues en contexte multilingue.* Thèse de nouveau doctorat, Spécialité Informatique, Université Joseph Fourier Grenoble I, 27 September 2001, 280 p.

Allan Melby et al. (1996) The Machine Readable Terminology Interchange Format (MARTIF), Putting Complexity in Perspective. Termnet News, vol.54/55, pp. 11-21.

Gilles Sérasset (1994) *Interlingual Lexical Organisation for Multilingual Lexical Databases in NADIA*. In Proc. COLING-94, Kyoto, 5-9 August 1994, M. Nagao ed. vol. 1/2 : pp. 278-282.

Gilles Serasset & Mathieu Mangeot-Lerebours (2001) *Papillon Lexical Database Project: Monolingual Dictionaries & Interlingual Links.* Proc. NLPRS'2001 The 6th Natural Language Processing Pacific Rim Symposium, Hitotsubashi Memorial Hall, National Center of Sciences, Tokyo, Japan, 27-30 November 2001, vol 1/1, pp. 119-125.

Bookmarks

GENETER modèle GENErique pour la TERminologie.

http://www.uhb.fr/Langues/Craie/balneo/demo_geneter.pl?langue=1

XLink 1.0 W3C Recommendation.

http://www.w3.org/TR/NOTE-xlink-req/

XML 1.0 eXtended Markup Language 1.0. W3C Recommendation.

http://www.w3.org/TR/REC-xml

XML Namespaces XML Namespaces. W3C Recommendation.

http://www.w3.org/TR/REC-xml-names

XML Schemas XML Schemas. W3C Recommendation.

http://www.w3.org/TR/xmlschema-0

XPath XPath Language. W3C Recommendation.

http://www.w3.org/TR/xpath

XPointer XML Pointer Language W3C Recommendation.

http://www.w3.org/TR/xpt

Annexs

Annex 1: XML Document Describing a Database

```
<database xsi:schemaLocation="http://www-clips.imag.fr/geta/services/dml
http://clips.imag.fr/geta/services/dml/dml.xsd"
name="GETA Lexical Database"
creation-date="22/10/99"
 owner="GETA">
  <partner-servers>
    <user-ref name="XRCE Analyser" href="xrce.xml"/>
  </partner-servers>
  <users>
    <user-ref name="Mathieu.Mangeot" href="mangeot.xml"/>
    <user-ref name="Mutsuko.Tomokiyo" href="tomokiyo.xml"/>
 </users>
  <groups>
    <group name="universe">
      <user-ref name="Mathieu.Mangeot"/>
      <user-ref name="Mutsuko.Tomokiyo"/>
    </group>
   <group name="lexicologists"><user-ref name="Mutsuko.Tomokiyo"/></group>
    <group name="administrators"><user-ref name="Mathieu.Mangeot"/></group>
  </groups>
  <dictionaries>
    <dict-ref name="FeM" href="FeM.xml"/>
    <dict-ref name="Papillon" href="papillon.xml"/>
  </dictionaries>
</database>
```

Annex 2: XML Document Describing a Dictionary

```
<dictionary
xsi:schemaLocation="http://clips.imag.fr/geta/services/dml
http://clips.imag.fr/geta/services/dml/dml.xsd"
category="multilingual"
creation-date="21/1/97 00:00:00"
encoding="ISO-8859-1"
format="rtf"
hw-number="192460"
installation-date="23/06/99 15:04:00"
fullname="dictionnaire français-anglais-malais"
name="FeM"
owner="GETA"
type="unidirectional"
version="1">
  <languages>
    <source-language lang="fra"/>
    <target-language lang="eng"/>
    <target-language lang="msa"/>
  </languages>
  <contents>general vocabulary in 3 languages</contents>
  <domain>general</domain>
  <bytes>9106261</bytes>
  <source>ML, YG, PL, Puteri, Kiki, CB, MA, Kim</source>
  <legal>all rights belong to ass. Champollion</legal>
  <cdm-elements>
    <headword delay="1s"/>
    <pronunciation delay="5s"/>
```

```
    <part-of-speech delay="5s"/>
    <translation lang="eng" delay="5s"/>
    <translation lang="msa" delay="5s"/>
    <corpus delay="10s"/>
  </cdm-elements>
  <administrators><user-ref name="Kim, ML"/></administrators>
  <volumes><volume-ref name="FeM" href="fem_fr_en_ms.xml"/></volumes>
</dictionary>
```

Annex 3: XML Document Describing a Volume

```
<volume
xsi:schemaLocation="http://clips.imag.fr/geta/services/dml
http://clips.imag.fr/geta/services/dml/dml.xsd"
 name="FeM_fr_en_ms"
source-language="fra">
  <entry>…</entry>
    ...
</volume>
```

Annex 4: XML Document Describing a User

```
<user
xsi:schemaLocation="http://www-clips.imag.fr/geta/services/dml
http://www-clips.imag.fr/geta/services/dml/dml.xsd"
name="Mathieu MANGEOT"
creation-date="22/10/2001">
  <login>Mathieu.Mangeot</login>
  <password>toto</password>
  <email>Mathieu.Mangeot@imag.fr</email>
  <profiles>
   <competences>
   <eng level="good">translation</eng>
      <fra level="mother tongue">phonetic, collocations, examples, grammar</fra>
      <jpn level="beginner"/>
      <spa level="good">translation</spa>
    </competences>
  <interests><interest lang="hun,jpn"/></interests>
  <activities>
   <activity dictionary="FeM">interface</activity>
   <activity dictionary="Papillon">administration</activity>
  </activities>
 </profiles>
 <credits>10</credits>
  <annotations href="mangeot-ann.xml"/>
 <contributions>
  <contribution source="French.xml" href="mangeot-cnt1.xsl"/>
  </contributions>
  <requests href="mangeot-req.xml"/>
  <xml-stylesheet type="text/css" href ="mangeot-sty.css"/>
  <groups>
    <group-ref name="universe"/>
    <group-ref name="administrators"/>
  </groups>
</user>
```

Annex 5: XML Document Describing a supplier API

```
<api type="supplier" category="consultation" name="JMDict_en-ja">
  <info>Dictionnaire japonais-anglais de Jim Breen</info>
  <url href="http://www.csse.monash.edu.au/cgi-bin/cgiwrap/jwb/wwwjdic"/>
```

```
  <protocol type="get"/>
 <delay min="1s" average="1s" max="2s" timeout="10s"/>
 <encoding input="UTF-8" output="EUC-JP"/>
  <format input="txt" output="html"/>
  <arguments>
    <element name="source-language">
      <complexType>
        <restriction base="string">
          <enumeration value="jpn"/>
          <enumeration value="eng"/>
        </restriction>
      </complexType>
    </element>
    <element name="headword" type="string"/>
    <element name="regex" type="boolean"/>
  </arguments>
  <result><element name="output" type="string"/></result>
</api>
```

Annex 6: XML Document Describing a client API

```
<api type="client" category="consultation" name="getabase">
  <info>API de consultation de la base lexicale du GETA</info>
  <url href="http://www-clips.imag.fr/cgi-bin/geta/dicoweb
mailto:dicoweb@imag.fr
telnet://www-clips.imag.fr:2628"/>
  <protocol type="post get mailto DICT" login="anonymous"/>
 <encoding input="ASCII ISO-8859-1 UTF-8" output="UTF-8"/>
 <format input="txt xml" output="xml html txt"/>
  <arguments>
    <element name="name" type="string"/>
    <element name="source-language" type="lang"/>
    <element name="word-order" type="string"/>
    <element name="cdm-elements" type="string"/>
    <element name="context" type="positiveInteger"/>
    <element name="input" type="string"/>
  </arguments>
  <result>
    <element name="output">
      <complexType>
        <sequence><element name="article" type="articleType"/></sequence>
      </complexType>
    </element>
  </result>
</api>
```

TOWARDS A GENERIC ARCHITECTURE FOR LEXICON MANAGEMENT

Cristina Vertan
Walther von Hahn,
University of Hamburg, Natural Language Systems Department
Vogt-Kölln-Straße 30, 22527 Hamburg, Germany
cri@nats.informatik.uni-hamburg.de, vhahn@nats.informatik.uni-hamburg.de

Abstract

In this paper we propose an architecture for a lexicon management tool MANAGELEX. This tool aims at a general environment for reading, updating and combining lexicons in different formats. The starting point is the already existing lexicon models MULTILEX and GENELEX. Each functionality (reading, updating and combining) is based on a corresponding model, which can be configured and maintained coherently.

1. INTRODUCTION

A large amount of lexical resources was developed during the last 15 years. Unfortunately, in the absence of a standard each application produced and used its own lexicon in a specific format and a specific model, according to particularities of language, system functionality and available physical resources. Reusable lexical resources, however, could noticeably reduce the cost of development of NLP applications. Moreover, during research projects, lexicon requirements may change over the run time of the project, and maintaining a suitable lexicon is expensive and time-intensive work.

The problem of standardization appeared as an absolutely and urgent necessity, and several projects were carried out in this sense (v.Hahn 2000). The task is quite difficult because it implies at least two components : standardization of the format and standardization of the model. Moreover, these two components are not completely independent. For the former it is general agreed today, that the starting point is a SGML –based format. Several SGML-lexicon standard formats were already proposed (EAGLES, OLIF, SALT) (Lieske & al. 2001, Melby 1999). It is, however, necessary that we have not only a standard set of tags but also a standard model of a lexicon representation. As a result of this insights, several projects tried to develop a standard and general model for lexicons. The most well-known formalisms after this phase are GeneLex and Multilex.

2. STANDARD LEXICON MODELS. STATE OF ART

Although having many architectural features in common, Genelex is abstracted basically from a French monolingual lexical model, whereas the Multilex architecture is genuinely designed as a multilingual language-independent general structure, trying to include all language specific models (EAGLES 1996). At least, as quoted in one of the final reports (Praprotté & al. 1993), Multilex *"is based on a consideration of the following languages: English, German, French, Spanish and Italian, and to lesser degrees Dutch and Greek"*. Compared to the multitude of (at least) European languages we observe that the Slavonic family was not taken into consideration, and also a lot of other languages which bring in new linguistic features (for example Romanian, although it belongs to the Latin languages, it has several important characteristics, due to the Slavonic influence).

The MULTILEX architecture presented a generic model for a lexical entry, which can be used as a starting point for further developments. However MULTILEX, as other similar projects *"imposes constraints on the linguistic level. Each of these projects imposes its own notion of 'lexical unit' (lemma, word-sense, concept) and its own logical structure (Typed Feature Structures, Entity-relationship model, automata, trees,...)"* (Sérasset 1996).

With these constraints, a user at the moment cannot use the same system to manipulate two lexicons coming from different places. Some steps in this direction were done in MULTILEX, which originally proposed the development of tools to convert lexicons into MULTILEX format. The proposal was not further developed because, quoting the same final report (Praprotté & al. 1993) *"copyright problems, problems in converting and correcting dictionary data, a lack of consistency in the data"* made this proposal unreachable.

Much lexical work from completed projects cannot be used in follow-up projects because of one of the following reasons:

- The lexicons were produced with the help of systems that are not any longer maintained; thus nobody can provide an export facility.
- In some cases, lexicon definitions contain procedural elements, which cannot be used without the hosting system,
- Lexicons may contain too rich features, which are too expensive to remove from the files.
- Experimental lexicons may be inconsistent or contain entries with different granularity,
- Lexicons may be stored in a data base, whereas others are plain files and the export formats do not match,
- Lexicons differ in their linguistic classes, i.e., there is a more-to-more mapping between feature classes.

From another point of view the use of a specific format (for example MULTILEX) means to adapt a posteriori other systems' processes to read and work which such external formats. This is usually quite cost-expensive.

The situation is much more critical for small languages, and languages from Central and East Europe, for which

lexical resources were developed quite ad hoc as they were needed for a certain project.

Although a lot of resources after a few years may be linguistically and technically outdated, about 60% of a dictionary with approx. 80 000 entries comprises the lexical core of very high and rather high frequency words, which remain stable in their syntactic and semantic properties over a long period of time. The other part (especially terminology) from time to time must undergo revision, updating or even replacements.

3. MANAGELEX A GENERIC LEXICON MANAGEMENT MODEL

Following the above considerations, we assume that for a rather long time from now, NLP applications will still have to deal with manipulations of non-standard lexical resources.

However, this is only possible with rather general lexical management tools for acquisition, comparison, manipulation and validation of lexicons, based on several abstract models.

In this section we propose a new architecture for a lexicon management tool (MANAGELEX), a tool, which is able to read, convert and combine lexicons, independent of their format, language or system requirements.
The general architecture of such a system includes (as shown in figure 1) 3 levels of abstraction (which follow the ANSI(1999) data modeling specifications): the meta model level, the model level and the real world level.

- The real world level identifies real (present), distinct objects, their concrete features, and the actual relation among them. In figure 1 this corresponds to the encoded lexicons (DocA, DocB) and their structure (StructA, StructB)

- The model level groups real world objects and present features into object and attribute classes and recognizes possible relationships among object classes. On this level our architecture has 3 tools:
  - A tool for reading and updating a lexicon (acquisition and editing tool),
  - a tool for encoding and decoding (encoding / decoding tool) and
  - a tool for mapping two lexicons, possibly with different structure (mapping tool)

- The meta model level, classifies types of elements appearing on the model level and the abstract relations among them, situation independent. Accordingly, we propose

  - A generic lexicon model (LexMod) which provides a rather rich model of possible lexical information. Here, every linguistic feature, with their possible values which may occur in a set of languages (at least European) are specified (MULTILEX together with the MILE (Calzolari & al. 2001) model (defined in the frame of the ISLE project) are a good starting point). A flexible formal specification will be provided for this model. The model will also allow for new categories, joining as well as splitting of existing categories.

  - A generic encoding model (Encod), which specifies the way of combining the linguistic information in a specific entry and lexicon structure. The model should also include options for encoding files in the new generally agreed SGML-standards as OLIF or SALT (Lieske & al. 2001; Melby 1999).

  - A mapping model (MAP), that specifies modalities of combining two lexicons and takes into account problems like mutual gaps and complex categories.

Given this architecture, we now explain the functionality of the envisaged system in three situations:

1. Building / updating a lexicon.

Input: Lexicon definition from LexMod, Encoding Model Encod,
Output: Lexicon interface, lexicon file

The operation is mainly performed by the acquisition/editing tool. The interface of this tool is built automatically according to the characteristics selected from LexMod for this particular lexicon. The output of this tool is a data structure recording the structure of the lexicon LexA. The encoding / Decoding Tool uses this data structure and the Encoding module and produces and encoded lexicon DocA.

2. Reading a lexicon.
Input: Lexicon file, Encoding Model Encod,
Output: -

This operation requires first the identification of the encoding and the generation of the corresponding linguistic structure (StructB). Responsible for all these is the encoding tool

3. Join of two lexicons (LexA and LexB)
Input: General Lexicon definitions from LexMod, lexicon definitions from StructA and StructB, mapping models MAP
Mapping models MAP
Output: Lexicon file

This is the most challenging operation. The mapping tool has to use not only the structure of the two lexicons (StructA and StructB) and the mapping model (MAP) but also the generic lexicon model (LexMod). This is required for example in case of different names for the same linguistic feature. The resulting structure contains data consistent with both lexicons. Furthermore a new lexicon can be encoded as described above.

4. CONCLUSIONS

In this paper we described a model of a possible lexicon management tool, which can deal with frequent problems in lexicon acquisition / maintenance. The presented architecture is still in prototyping phase. We envisage to develop it in the frame of an European project. How ever for the moment we will take into account the European languages. Extensions to other language should be possible one the system reaches a stable version. The system is not intended to replace the actual already defined standards, but to supply the use and reuse of the already developed non-standard lexical resources

REFERENCES

ANSI-American National Standard Institute(1999), Standard X3. 138-1988, *Information Resource Dictionary System (IRDS)*

Calzolari, N. and A. Lenci and A. Zampolli and N. Bel and M. Villegas M. and G. Thurmair G., (2001) "The ISLE in the Ocean – standards for Multilingual Lexicons (with an Eye to Machine Translation*)", Proceedings of MT Summit VIII, Santiago de Compostella, 2001*

EAGLES (1996) "Input to the EAGLES architecture work: survey of MULTILEX", *http://www.ilc.pi.cnr.it/EAGLES96/lexarch/node4.html*

v.Hahn, W (2000), "Standards in Natural Language Processing – New Steps in Language Engineering", in *Standards in Information technology S. Nedevschi and K. Pusztai (Eds.*), Casa Cartii de Stiinta, Cluj.

v.Hahn, W. (1999), "Metamodelling of Lexical Acquisition Tools", *Proceedings of EUROLAN '99*, Iasi.

Lieske Ch. and S. McCormick and G. Thurmair (2001), "The Open Lexicon Interchange Format (OLIF) comes of Age", *Proceedings of MT Summit VIII, Santiago de Compostella*

Melby, A. K. (1999), "SALT: Standards-based Access service to multilingual Lexicons and Terminologies", *http:// www.ttt.org*

Paprotté, W. and F. Schumacher(1993), "MULTILEX – final Report WP 9: MLEXd", *Report MWP 8 – MS*

Sérasset G.(1996), "Recent Trends of Electronic Dictionary. Research and Development in Europe", *Report GETA-IMAG, CNRS, Grenoble,*

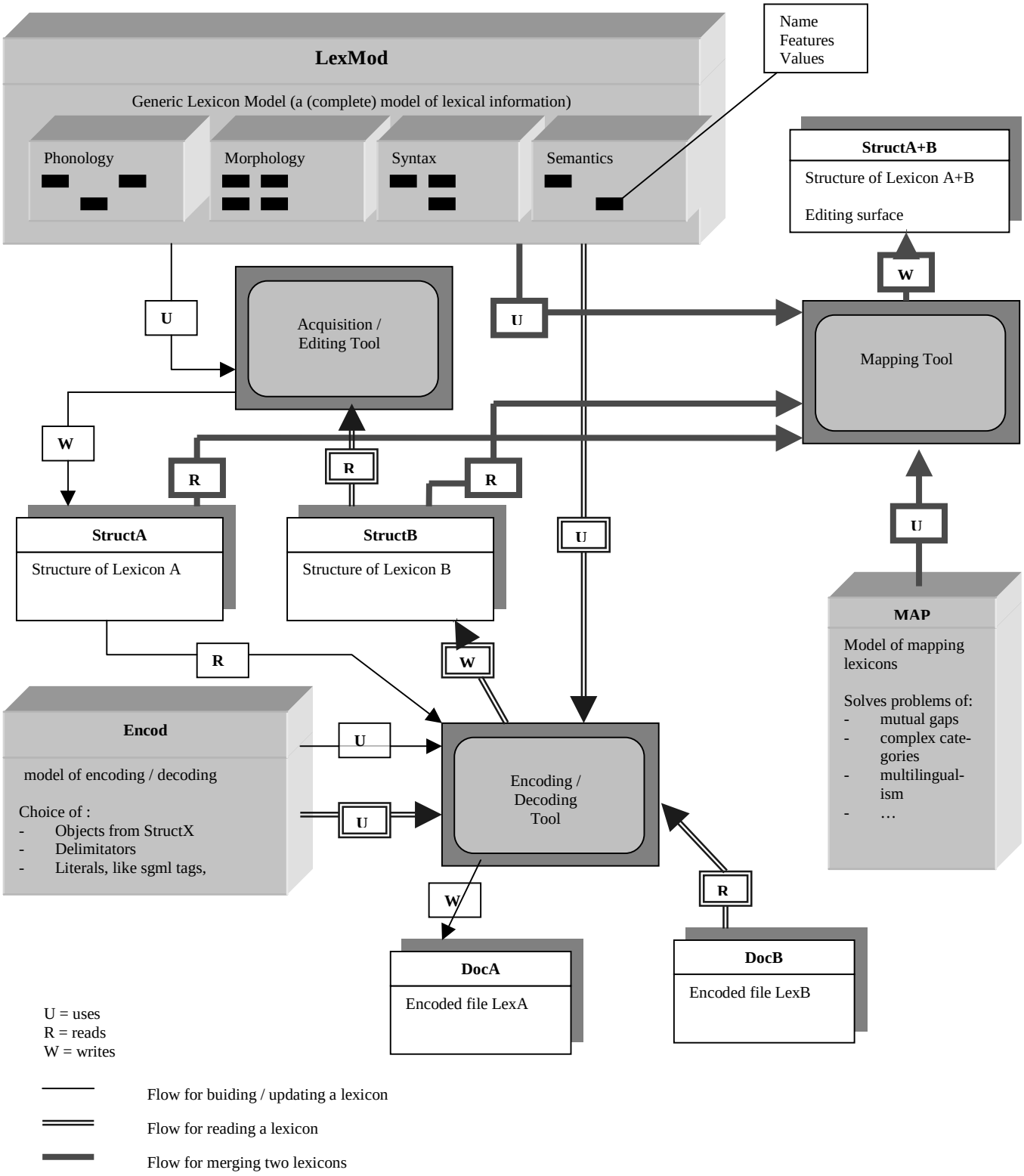


Figure 1: MANAGELEX: components and Workflow

Management of Language Resources using Metadata

P. Wittenburg, Daan Broeder

Max-Planck-Institute for Psycholinguistics
peter.wittenburg@mpi.nl

Abstract

Technology development allows many more researchers than before to create language resources especially with multimedia extensions. This creates a resource management problem that exceeds the boundaries of established resource centers. Metadata environments such as the one proposed by IMDI that offer a metadata set and also tools to operate on them have a strong potential to help the individual researcher to carry out his resource management tasks. In addition, it allows him to easily integrate his resources into a large distributed domain of resources. The work at the Max-Planck-Institute for Psycholinguistics to establish a large multimedia language corpus helped to understand the needs and requirements. Due to this experience the IMDI environment has reached a state of maturity, but still some important features have to be added.

1. Introduction

Researchers and developers in the area of language resources are faced with four very dominant trends in the recent years: (1) The number and complexity of language resources stored in digital archives is growing fast, (2) there is an increasing acceptance of the need to improve the availability of the resources, (3) the Internet now connects many archives storing such resources and this asks for interoperability and (4) for many language resources need to be stored in archives for a large period of time due to economical and ethical reasons.

An impression about this explosion of resources can be given by the example of the multimedia/multimodal corpus at the Max-Planck-Institute for Psycholinguistics where every year around 40 researchers carry out field trips, do extensive recording of communicative acts and later annotate the digitized audio and video material on many interrelated tiers. The institute now has already more than 7000 annotated sessions - the basic linguistic unit of analysis - and we foresee a continuous increase. It was usual that researchers managing their resources with individually designed Excel-Sheets eventually were not able to keep control of them and that the institute effectively lost all access to resources when a researcher left. Thus the individual researcher as well as the institute was both faced with a resource management problem. It is known that in other research centers, universities and also in industry similar situations occur.

The increase of the amount of resources was paralleled by an increase in the variety and complexity of formats and description methods. Moving from purely textual to multimedia resources with multimodal annotations caused this. Media can include not only several audio and video tracks, but also increasingly often other information such as for example from eye trackers, data gloves and brain image recorders.

In many areas resources were seen as the private capital of a researcher or a specific project that served only to investigate a limited number of research questions. Therefore, the need to make resources available for other research was not seen. However, researchers now understand the potential of modern technology to immediately access the raw material, which enables for example re-coding, or incremental annotation procedures that can be part of collaborations. These opportunities increase the individual researchers willingness to share his resources and to invest time to create publicly available descriptions. We clearly recognize a trend towards making the resources themselves available via the Internet or at least indicating what resources exist by creating structured descriptions available on the Internet.

The usage of the Internet demands for interoperability on various levels. Therefore new technologies devoted to the special requirements of the Internet such as RDF (Resource Description Framework), XML and UNICODE are have been developed to improve the exchange and re-usage of data. The usage of open standards is even more important when repositories of language resources have to support long archive periods. The Internet also adds another dimension of complexity since people want to create distributed repositories where the resources of a corpus can be scattered over different locations, nevertheless requiring transparent access to them.

Summarizing we can say that a much broader group of researchers besides the experts who have always handled expensive resources are now involved. They are managing larger amounts of more complex structured resources, making them available in standardized formats and descriptions via the Internet. Now that resource creation has become much more easy many individual researchers are also coping with resource management problems pushing the management task beyond the experts at large data centers.

2. Resource Management

The increased relevance of resource management can best be seen in the document domain by the emergence of various sorts of commercial Content Management Systems. It is widely understood that only improved management concepts will allow us to prevent a chaotic situation where we will have an increasing amount of data on our storage devices, but don't know about them nor know how to access them.

We can identify at least four different groups of people involved in resource management each one with their own views: (1) the computer system specialists have to be able to manage data on a physical level. They allocate physical resources, define structures in file systems and take care of redundant copies for secure data storage. (2) The producer of resources wants to integrate his resources into the repository in an easy way and describe them easy and correctly to facilitate retrieval. (3) The user wants to deal with data on a domain-oriented level, i.e. a level where the

well-established concepts and terminology of a domain are used. He is not interested in file system details. This view includes distributed scenarios where the user wants to combine resources from different institutions without having to know where exactly the resources reside. Often the producer is himself a user. (4) The archive manager acts as an interface between system specialists and producers and also prefers to manage data at the level of domain concepts. At least he has to know how the system managers handle the resources since he has to draw the links between logical and physical structure and influence for example the policies for protecting the data. In many cases the producer/user is also the archive manager, since there is no support stuff. Management has to consider all views.

The following is a non exhaustive list of points to be addressed by modern resource management (resource discovery is in general seen as being a component of resource management, but in this paper we will mention it, but not focus on it).

- How to store resources such that they can survive for many years independent from technology changes.
- How to protect resources against unauthorised access
- How to create personalized views on resource repositories to facilitate easy and optimised navigation
- How to offer easy and immediate access to resources after access is approved?
- How can descriptions of sets of resources be modified easily?
- How to easily integrate new resources into the distributed resource repository?
- How to keep track of old versions?
- How to make such a management scheme available to interested parties.
- How to easily move groups of resources to other locations transparent to the user/producer?
- How to achieve hardware and operating system independent operation within the resource domain?
- How to easily integrate different data types that belong together and allow access while hiding the complexity?
- How to inform people about the existence of a resource and its major characteristics?
- How to easily discover resources in a distributed scenario from a conceptual perspective?

In this paper we will focus on the resource manager and user views. This although many important problems such as for example the problems of long-term archiving of digital media are not at all solved.

3. Pillars of Management

As already indicated, industry delivers a wide range of software solutions that are meant to cover documents of all sorts. In this paper we will not discuss Document Management Systems although they may deliver much functionality, but focus on the key pillars of open distributed solutions aimed at our specific environment and data types.

3.1. Standards

Open standards are very important to achieve interoperability, to build up long-term archives and to produce long-term available tools. Especially in the domain of computer-based language resources, however, we are faced with an extremely dynamical situation. This means we are confronted with a multitude of standards making many people turn over to use the word “best practice guidelines” instead. For multimedia resources for example we are confronted with a long list of media compression methods (MPEG1/2/4, Cinepak, Sorensen, MP3, ATRAC etc) all emerging within the last decade. Each having its advantages and disadvantages dependent on the field of application. For an archive one has to decide about major backend standards (such as MPEG2) which allows creating other representations for specific applications on the fly.

Referring to the earlier questions we need a couple of standards. We claim that many of the management problems can be solved with the help of establishing a suitable metadata environment existing of a metadata element set and appropriate tools. Tools themselves are not subject of standardization per se, since it is good to have competing solutions. With respect to the metadata set, however, we need agreements on various levels. The metadata elements are the dimensions of how to characterize a resource and it is clear that each choice for a set of dimensions limit the expressiveness for other groups of users. Therefore, we can expect that there will be different sets of dimension to describe multimedia/multimodal language resources. Important for the community is that we have open accessible definitions of the elements such that schemes can refer to them. They should be described as Data Categories if this will be the common practice for terminology repositories.

In addition, in the case of non-orthogonal spaces as the one we need to describe, these dimensions can only be defined appropriately by specifying suitable controlled vocabularies. They are the values that a specific dimension can take. Also these controlled vocabularies have to be openly accessible and should be defined in the same way. Both elements and their controlled vocabularies, have to be known exactly to achieve interoperability. Of course, it makes sense to use just one controlled vocabulary for example for language codes, but also here we are faced with different (quasi) standards such as ISO 639-2, the Ethnologue list from SIL[1] [1,2] and the various lists handled by specific projects. Also here we must accept that different vocabularies will exist.

Consequently, we are faced with mapping problems on different levels. RDF will be the primary language to try and bring all the different pieces of the mosaic together. This problem has not been tackled yet with the exception of a few cases such as in the Harmony project and in the mapping proposal from IMDI[2] to DC[3]/OLAC[4]. MPEG7[5] categories were mapped on Dublin Core categories in a very restricted way and the element relations are described

[1] Summer Institute of Linguistics
[2] ISLE Metadata Initiative
[3] Dublin Core Metadata Initiative
[4] Open Language Archives Community
[5] MPEG7 is the standard for media annotation within the family of MPEG standards in the film and media industry

with the help of the RDF formalism. Such a formal framework has not yet described the IMDI to OLAC mapping. At the moment we don't know which expressional power the community will need to accomplish the big task to create such a mapping for the language resource domain. The emergence of DAML/OIL [3] indicates, however, that RDF itself will probably not be sufficient.

It is assumed here without further comment that XML is our common language, i.e. all definitions and frameworks to be used should be based on XML.

3.2. Metadata Descriptions

The usage of metadata descriptions for improving the management of documents is not a new concept. Librarians are used to describe their documents with cards since many years. Linguists and speech engineers were used to describe characteristics of their resources and put these in file headers - mostly project specific formats. The community learned a lot from the TEI [6] work about standards for resource headers (later adopted by the CES[7]) and it is still used as a reference to look at. Also in some projects such as CGN[8] the TEI recommendations were followed to a certain extent.

TEI is a comparatively exhaustive descriptor set meant to describe the characteristics and structure of a resource. Newly developed metadata sets do not want to describe the resource in a too great detail, but address the problem of easy discovery primarily, i.e. a resource would be described sufficiently well, if a user manages to find it. Metadata sets such as DC, OLAC and IMDI follow this approach. DC tries to address the discovery problem with 15 sloppily defined categories ordered in a flat structure. In doing so DC allows the user to describe resources about steam engines as well as resources about Sign Language both on a very general level. For many DC categories it is not clear how they can be applied to different domains, therefore refinements are defined as was done by the OLAC initiative. The "DC:Type" element that defines the resource type is refined by the characteristic "CPU" to describe the type of CPU a NLP tool can run on. The semantics of such an element are stretched extremely.

MPEG7 and IMDI followed another approach since they started with studying the domain specific requirements. For MPEG7 it is essentially the production process of movies that has to be covered to later be able to retrieve relevant segments that are covered by the metadata set in addition to the ordinary elements such as "Creator". The basis of IMDI was an extensive survey of the different ways in which linguistic resources in all their variety have been described. Often this was done in the form of a proprietary "file-header" that contained metadata information about the annotation as a whole such as for instance the CHAT file format [4]. CES (being TEI compliant with respect to corpora) suggestions were applied were useful for discovery, however, we have not found sufficient support for other types of linguistic data than text. TEI/CES also mixes metadata and content in the same way as MPEG7. IMDI has favored a physical separation of metadata and content allowing uncomplicated protection schemes which is important for some groups of users. It also allows separate management of resources and metadata, usefull because the integration of legacy data formats has to be supported.

[6] Text Encoding Initiative
[7] Corpus Encoding Standard
[8] Spoken Dutch Corpus Project

3.3. IMDI

3.3.1. Session Concept

The IMDI set was especially targeted at multimodal/multimedia resources and their inherent complexity, i.e. basis is in general the existence of media recordings. This led to the development of the "Session" concept. For linguists a session is defined as the basic unit of linguistic analysis and covers a coherent type of linguistic action or performance. From a corpus organization point a session is the leave in the tree. A session is in general associated with a bundle of tightly related resources: a video recording of a native speaker, a set of pictures of that persons house, some field notes about this scene and afterwards some multimodal annotations. The IMDI definition of the term "session" covers this bundling from an access and management point of view.

In DC one would have to use the "DC:Related" element to describe the relation between these resources that is associated with much overhead. This was described in more detail in the IMDI-OLAC mapping document [5].

From a management point of view the session concept makes sense since accessing or extracting subcorpora implies accessing resp. copying of complete sets of related information.

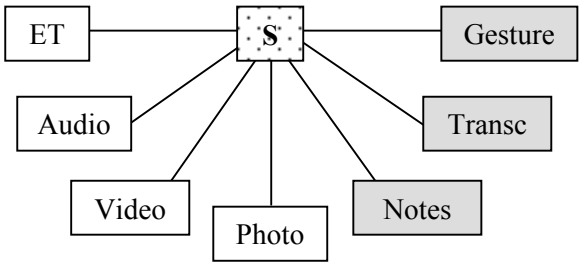


Figure 1 shows a typical session with its related resources all referring to the same linguistic event. It covers different types of recordings and different annotations.

In IMDI its the structured metadata set which describes this relation, i.e. there is only one metadata description (if the user decides to do it that way) with different sub-blocks describing the characteristics of the individual components. This way allows a user to ask questions such as "give me all resources which have eye movement recordings and a phonetic transcription of what was spoken"

3.3.2. Browsable Domain

Next to the "Session" concept, IMDI introduced the idea of structuring corpora in a conceptual space by having hierarchies of (sub-) corpora where description nodes representing a certain level of abstraction with respect to other (sub-) corpus nodes culminating eventually in pointers to session nodes (see figure 2). Each level represents a certain abstraction layer that is meaningful to the resource manager or user.

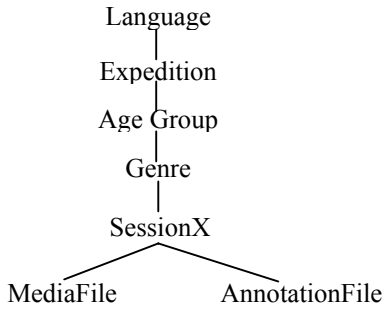


Figure 2 shows a typical hierarchy from field linguistics

Since corpus nodes create logical structures several parallel hierarchies can be created to structure the same (sub-)corpus and to express different interests of users. This allows each user to establish his own preferred view on the distributed resource domain and by also using bookmarks to create his own conceptual space (see figure 3). These parallel hierarchies can also be used to support versioning. Of course, there is no reason for the user to not create cross-references. For management purposes such cross-references are of course difficult to handle, i.e. the resource managers preferably would work with just the canonical tree.

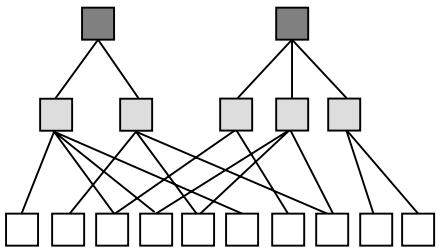


Figure 3 shows two user defined hierarchies referring to the same set of session nodes that are at the bottom level. One view could make a sex distinction, another one by age groups.

The mechanism by which the (sub-) corpora nodes refer to each other is to use URL's. This has the advantage to support distributed corpora frameworks and create a unique namespace for all resources.

3.3.3. Data Type Integration

Such a browsable domain as indicated is of course very useful for integrating various data types that we find in complete corpora. We already described the integration on the session level. For many data types however it only makes sense to associate them with higher nodes in a corpus tree. Such a node represents an abstraction with respect to a number of metadata elements (for example sharing the same language). Lexica can be related to a sub-corpus associated with a language or a set of recordings for a language (lexicon of a 3 year old child). Field notes and comments about dialect variants in general can appear on all levels of a corpus. In general many of these data types do not have any definite structure, but are just prose texts in some general format such as DOC, HTML or PDF. Corpus management has to provide mechanisms to include such descriptions in a flexible way. IMDI allows the resource manager to do so, but of course, will exclude proprietary formats such as DOC.

3.3.4. Practical Considerations

A strong concern was and still is how one can enforce creators and managers to adhere to standards with all its consequences as described above. The stricter the rules are such as full adherence to the chosen controlled vocabulary of a certain element, the more sensitive these procedures will become. Although the IMDI type of operations are now in operation for 3 years we cannot claim that a "standard" such as IMDI for describing language resources will not undergo changes. In IMDI for example we expect changes with respect to the dimensions and vocabularies that describe the resource content.

It was found - and this experience is nothing new - that it is very important to support the creators and managers with professional tools. Within IMDI it was always tried to have a balance between the development of the metadata set and an editor that supports the creation of IMDI descriptions. The IMDI editor now supports

- All metadata elements including their controlled vocabularies in a dynamic way, i.e. if the definition in the repositories change the editor will adapt its representations
- Sub-blocks which allow the user to save and reuse reoccurring information such as participant or project information

Version changes in the metadata set can of course lead to severe problems for corpus management and metadata usage. There efficient tools are of the greatest importance to modify all whole sets of existing metadata descriptions. Currently, a script allows the resource managers to change the values of the elements for a whole set of metadata descriptions. Of course, such operations are very sensitive and such a script may not be given to the general user. The intention is to include such an option in the editor such that all changes are conforming to the actual IMDI definitions.

The browser offers the same feature as the editor in so far that it also uses the actual vocabulary definitions from the repository. Further, the browser offers the following management relevant features:

- A user can create new (private) nodes and therefore define his own view on a sub-corpus
- It is possible to start the editor from the browser environment to modify metadata descriptions
- It is possible for the users (managers) to associate tools with individual or bundles of resources such that when a (set of) useful resources was found immediately a tool can be started to operate on the resources.

Both tools will have to provide for version conversion in case they find metadata descriptions in an older format. They should not however work with old versions without forcing (if possible) an update.

In the future the editor has to be extended to be able to create formatted lists (Spreadsheet type) of the content of a range of metadata descriptions for easy check and input to for example statistic programs. This is a favorite view on metadata of many users. The user has to be able to select the elements he wants to see. One complication is

given through the fact that some elements can occur several times such as participants, i.e. the number of entries for the spreadsheet can only be computed by first reading all selected metadata descriptions.

3.3.5. Difference to Normal HTML Domains

Of course, the basic organization principles sound very familiar, since we use the same for designing web pages. Instead of creating XML based descriptions one could create HTML pages and include all information and data types as hyperlinks in the usual way. Some archives are operating this way. Metadata descriptions could be included in the headers of the HTML files to support element-based search.

The IMDI team did not choose for this way for the following major reasons:

- HTML is basically a way to describe how documents should be displayed and not to describe data structures.
- Using HTML would not have made sense without also using HTTPD servers and browsers. Otherwise HTML is just a much less powerful version of XML. The current HTML browsers however are not suited to perform all computation tasks required of a metadata browser such as making intelligent choices for tools to work on resources.
- We needed a format to transfer information. Tools should be able to interpret this information either to display parts of it or to offer the user a choice of tools to work on referenced resources.

4. Conclusions

Based on 3 years of experience with a multimedia/multimodal corpus which covers already more than 7000 metadata descriptions and a showcase application including sample corpora from 6 European institutions we can draw some conclusions.

1. All questions raised in chapter two are addressed by the IMDI environment with two exceptions: (1) Version handling of resources and metadata description schemes are not yet supported by the tools by the tools. (2) The tool for extracting complete sub-trees of a corpus is not yet available.
2. The need to apply the definitions and tools to such a big and heterogeneous corpus as for example the MPI corpus was a useful and necessary enterprise. It made us understand the underlying processes and requirements to establish an environment such as IMDI.
3. Corpus management was performed during the development phase of the IMDI environment. This meant that frequent updates of the metadata schema took place that required frequent transformation of the metadata files.
4. We now have an environment where it is comparatively easy to integrate or build up IMDI based archives that supports the creator, the user and especially the resource manager with suitable mechanisms and tools.
5. Since all definitions are open everyone can create his own set of tools to work on the metadata descriptions, i.e. improve the search engine or write another browser.
6. Using a file oriented framework for storing metadata only appears as an advantage when distributing or integrating small (personal) archives or making extractions of sub corpora on portable media for off-line use. It does however create confidence of the linguists that they can take their metadata descriptions with them on a floppy and are not dependent on server bound DBMS's.
7. Using metadata in a uniform, controlled and structured way is a new experience for our linguists. It did and still costs a large persuasion effort to have them input their metadata. It has only be since a short time that they themselves can reap the benefits by using for instance metadata search, since a critical mass is necessary and since the improvements for resource management had to become apparent.
8. The introduction of a complete and operational metadata environment was the first experience for the development team of this sort. Often the practical experience guided us in designing and improving the tools, since we did not foresee all aspects of efficient resource management beforehand.

Finally, it seems to be appropriate to add a statement about future perspectives. We see metadata for language resources still in its beginning phase, since there are not so many resource repositories which already created the appropriate files. Especially there are only few attempts to do resource management with the help of metadata environments. We have shown their great potential but also the difficulties involved. Especially the inclusion of metadata element and vocabulary definitions in open repositories and the formulation of their relations with the help of Semantic Web compliant mechanisms such as RDF will motivate more groups to contribute and participate. Interoperability between different metadata sets will also be facilitated by applying these agreed standards.

The soon to be started INTERA project is aiming to realise and work at the above mentioned points.

[1] [ISO639-2]
Codes for the representation of names of languages - part 2: alpha-3 code, International Organization for Standardization (ISO), 1998. http://lcweb.loc.gov/standards/iso639-2/langhome.html

[2] Ethnologue language name index http://www.sil.org/ethnologue/names/

[3] DAML/OIL: http://www.daml.org

[4] Childes: http://childes.psy.cmu.edu

[5] IMDI-OLAC-Mapping: http://www.mpi.nl/ISLE

Towards Multimodal Content Representation

Harry Bunt*, Laurent Romary&

Computational Linguistics and AI, Tilburg University
P.O. Box 90153, 5000 LE Tilburg, The Netherlands
Harry.Bunt@uvt.nl
&LORIA, University de Nancy
B.P. 239, 54506 Vandoeuvre-les-Nancy, France
Laurent.Romary@loria,fr

1. Introduction

Multimodal interfaces, combining the use of speech, graphics, gestures, and facial expressions in input and output, promise to provide new possibilities to deal with information in more effective and efficient ways, supporting for instance:

- the understanding of possibly imprecise, partial or ambiguous multimodal input;
- the generation of coordinated, cohesive, and coherent multimodal presentations;
- the management of multimodal interaction (e.g., task completion, adapting the interface, error prevention) by representing and exploiting models of the user, the domain, the task, the interactive context, and the media (e.g. text, audio, video).

An intelligent multimodal interface requires a number of functionalities concerning media input processing and output rendering, deeper analysis and synthesis drawing at least upon underlying models of media and modalities (language, gesture, facial expression of user or animated agent), fusion and coordination of multimodal input and output at a semantic level, interpretation of multimodal input within the current state of the interaction and the context, and reasoning about and planning of multimodal messages. This implies an architecture with many components and interfaces; a reference architecture of an intelligent multimodal dialogue system was established at the workshop `Coordination and Fusion in Multimodal Interaction' in Dagstuhl, Germany, November 2001 (see Bunt, Kipp, Maybury and Wahlster, forthcoming, and http://www.dfki.de/~wahlster/Dagstuhl_Multi_Modality). The communication between many of the components in a multimodal interactive system rely upon an enabling syntax, semantics and pragmatics. A multimodal meaning representation plays central stage in such a system, supporting both interpretation and generation. Such a representation should support any kind of multimodal input and output, and should, in order to be useful in a field which is still developing, be sufficiently open to support a range of theories and approaches to multimodal communication.

The present document is intended to support the discussion on multimodal content representation, its possible objectives and basic constraints, and how the definition of a generic representation framework for multimodal content representation may be approached. It takes into account the results of the Dagstuhl workshop, in particular those of the informal working group on multimodal meaning representation that was active during the workshop (see http://www.dfki.de/~wahlster/Dagstuhl_Multi_Modality, Working Group 4).

2. Scope

To delineate the task of formulating objectives, constraints and components of multimodal meaning representation, we must first have a shared understanding of what is meant by *meaning* in multimodal interaction. We propose to define the meaning of a multimodal `utterance' as the specification of how the interpretation of the `utterance' by an understanding system should change the system's information state (taken in a broad sense of the term, including domain model, discourse model, user model, task model, and maybe more - see e.g. Bunt, 2000). While formulated with reference to input interpretation only, this definition can also be related to the generation of multimodal outputs, by assuming that an output is generated by the system in order to have an effect on the user through the interpretation of that output by the user. (The generation of appropriate outputs thus depends on the system having an adequate model of what its outputs may mean to the user – which is exactly as it should be.)

A multimodal meaning representation should support the fusion of multimodal inputs and the fission of multimodal outputs at a semantic level, representing the combined and integrated semantic contributions of the different modalities. The interpretation of a multimodal input, such as a spoken utterance combined with a gesture and a certain facial expression, will often have stages of modality-specific processing, resulting in representations of the semantic content of the interactive behavior in each of the separate modalities involved. Other stages of interpretation combine and integrate these representations, and take contextual information into account, such as information from the domain model, the discourse model and the user model. A multimodal meaning representation language should support each of these stages of interpretation, as well as the various stages of multimodal

output generation. Since we are considering inputs and outputs from a semantic point of view, the representation of lower-level modality-specific aspects of interactive behavior, like syntactic linguistic information or morphological properties of gestures is not a primary aim, but some such information may percolate as features associated with a meaning representation, especially at intermediate stages of interpretation, where their relevance for semantic interpretation may not have been fully exploited. At the other end of interpretation, where understanding is rooted in information structures like domain models and ontologies, a multimodal meaning representation language should support the connection with frameworks for defining ontologies and specifying domain models, such as DAML + OIL.

While supporting the linking up of meaning epresentations with ontologies and `low-level' modality-specific information, the design of multimodal meaning representations is to be clearly distinguished from the design of domain model representations, linguistic morphosyntactic representations, representations of facial expressions, etc., which do not fall within this scope. Also, meaning representations should not represent the underlying processes by which they are constructed and manipulated, although it may be important that they are `annotated' with administrative information relating to their processing, such as time stamps.

3. Objectives

The main objective of defining multimodal meaning representations is to provide a fundamental interface format to represent a system's understanding of multimodal user inputs, and to represent meanings that the system will express as multimodal outputs to the user. This interface format should thus be adequate for representing the end result of multimodal input interpretation, and for representing the semantic content that the system will present to the user in multimodal form. It should therefore allow dialogue management, planning and reasoning modules to operate on these representations. In order to be useful for this purpose, this interface format should support the interfaces of these as well as other modules that form part of the system, and thus be adequate not only for representing the end result of semantic interpretation but also intermediate results. Something similar holds for generation. This is a second objective that follows almost immediately from the first.

Another objective in defining a well-defined representational framework for multimodal communicative acts is to allow the specification and comparison of existing application-specific representations (e.g. the M3L representation used in the SmartKom project) and the definition of new ones, while ensuring a level of interoperability between these.

Finally, the specification of a multimodal meaning representation should also be useful for the definition of annotation schemes of multimodal semantic content.

4. Basic Contrains

Given the main objective of defining meaning representations, the first and foremost basic requirements of a semantic representation framework are those that we may call `expressive' and `semantic' adequacy:

- *Expressive adequacy:* the framework should be expressive enough to correctly represent the meanings of multimodal communicative acts;
- *Semantic adequacy:* the representation structures should themselves have a formal semantics, i.e., their definition should provide a rigorous basis for reasoning (whether deductive, statistical, in the form of plan operators, or otherwise).

The second objective, of providing interface formats within a multimodal dialogue system architecture, means that `incremental' construction should be supported of intermediate and partial representations, leading up to a final representation or, if the construction of a final representation does not succeed, leading to negative feedback or another appropriate system action. This implies three further basic constraints:

- *Incrementality,* in the sense of supporting various stages of multimodal input interpretation, as well as of multimodal output generation, allowing both early and late fusion and fission;
- *Uniformity:* to make incremental processing feasible, where possible the representation of various types of input and output should be uniform in the sense of using the same kinds of building blocks and the same ways in which complex structures can be composed of these building blocks.
- *Underspecification and Partiality*: to support the representation of partial and intermediate results of semantic interpretation, the framework should allow meaning representations which are underspecified in various ways, and which capture unresolved ambiguities.

Finally, the representational framework should take into account that the design of multimodal human-computer dialogue systems is a developing area in which new research results and new technologies may bring new challenges and new approaches for the representation of multimodal meanings. This means that the representational framework should satisfy the following two constraints:

- *Openness:* the framework should not depend on a single, particular theory of meaning or meaning representation, but should invite contributions from different semantic theories and approaches to meaning representation;
- *Extensibilty*. The framework should be compatible with alternative methods for designing representation schemas (like XML), rather than support only a single specific schema.

5. Methodology

As a first step in the direction of defining a generic multimodal semantic representation form, we have to establish some basic concepts and corresponding terminology.

First, the action-based concept of meaning mentioned above, applicable to multimodal inputs in an interactive situation, means that the meaning of a multimodal `utterance' has two components: one that is often called `propositional' or `referential', and that is concerned with the entities that the utterance refers to and with their properties and relations that may be expressed in propositions, and a `functional' component that expresses a speaker's intention in producing the utterance: what effects does he want to achieve (using `speaker' in a broad, multimodal sense here)? This distinction is familiar from speech act theory, where the two components are called `propositional content' and `illocutionary force', and is also prevalent in other theories of language-based communication (see Bunt, 2000); it is often viewed as drawing a border line between semantics and pragmatics. In the analysis of multimodal interaction it is especially important to pay attention to both these aspects of meaning, since different modalities often contribute to each aspect in different ways; for instance, in spoken interaction the referential and propositional aspects of meaning are often expressed verbally, while gestures and facial expression contribute primarily to the functional aspects. The term `multimodal content' should not be confused with `propositional content', and should not make us forget that multimodal messages have meanings with functional aspects that are equally important as their propositional and referential aspects. In this document we use `multimodal content' as synonymous with `multimodal meaning', including functional aspects, and we use `semantic representation' as synonymous with `representation of meaning'.

A convenient term that has become popular in the literature on human-computer dialogue is *`dialogue act'*. This term is mostly used in an informal, intuitive way, or as a variant of `speech act; it has a formal definition in terms of the effects that a `speaker' intends to achieve through its understanding by the addressee (see Bunt, 2000), which makes it suitably precise for use in the analysis of the meaning of multimodal inputs and outputs. Without further going into definitions here, we will use the term `dialogue act' in the rest of this document. Definitions of other useful concepts can be found in Romary (2002).

As a second methodological step, we propose to distinguish the following three basic types of ingredients that would seem to go into any multimodal meaning representation framework. Each of these ingredients is discussed further in subsequent sections

1. *Basic components*: the basic constructs for building representations of the meaning of multimodal dialogue acts: types of building blocks and ways to connect them.
2. *General mechanisms*: representation techniques like substructure labeling and linking, that make the representations more compact and flexible.
3. *Contextual data categories*: types of administrative (meta-)data that do not, strictly speaking, contribute to the meanings of semantic representations, but that may nonetheless be relevant for their processing.

5.1 Basic Components

Initially, the following basic components can be identified to represent the general organization of any semantic structure:

1. temporal structures (*`events'*), to represent, for instance:
   - spoken utterances (input or output dialogue acts);
   - gestures (same);
   - noncommunicative action (like searching for information, making a calculation);
   - events, states, processes,.. in the discourse domain, representing meanings of verbs and possibly other linguistic expressions;
2. referential structures (*`participants'*), to represent, for instance:
   - the speaker of an input utterance, or the person performing a gesture;
   - the addressee of a system output dialogue act;
   - individuals and objects participating in a semantic event
3. *restrictions* on temporal and referential structures, to represent, for instance:
   - the type(s) of dialogue, act associated with an utterance;
   - a gesture type, assigned to a gesture token
4. dependency structures, representing *semantic relations* between temporal and/or referential structures, for instance:
   - participant roles (like SPEAKER, ADDRESSEE, AGENT, THEME, SOURCE, GOAL,..)
   - discourse/rhetorical relations
   - temporal relations.

It may be noted that linguistic semantic phenomena that have been studied extensively in relation to the needs of underspecific representation, such as quantification and modification, can also be represented with these basic components. For instance, a quantified statement like `Three men moved the piano' can be represented as a move-event involving a group of three men and a piano, where the collectiveness and the group size of the set of men that form the agent of the event are represented by means of restrictions on the event.

5.2 General Mechanisms

In addition to these basic components, certain general mechanisms are important to make meaning representations suitable for representing partial and underspecified meanings, to give the representations a more manageable form, and to relate them to external sources of information. Examples of such mechanisms are:

1. *substructure labeling*: assigning labels to subexpressions and allowing the use of these labels, instead of the substructures that they label, as arguments in other subexpressions;
2. *argument underspecification*: partial or underspecified representations can be constructed using labels in argument positions; restrictions on labels can represent limitations on the ways in which such variables can be instantiated by labels of substruct - ures elsewhere in the representation;
3. *restrictions on label values*: see previous mechanism. Alternatively, *disjunctions*, or *lists* of labels can be used to represent ambiguity or partiality;
4. *structure sharing*, as in typed feature structures, makes it possible to represent that a certain part of the representation plays more than one role, e.g. a participant may be both agent and theme in a semantic event, or may be the speaker of an utterance and the performer of a gesture, as well as the agent in a semantic event expressed by the multimodal dialogue act;
5. *linking to domain models* (types and instances) to anchor meaning representations in the domain of discourse;
6. *linking to lower levels*, such as syntactic structure, prosodic cues, gestural trajectories,.. is useful for tying a purely semantic representation to lower-level information that has given rise to it, and that may not yet have been fully interpreted.

5.3 Contextual Data Categories

Finally, meaning representations will need to be annotated with general categories of administrative information, both globally and also at the level of subexpressions, to capture certain information which is not found inside the elements of interactive behaviour, but which is potentially relevant for their interpretation and generation, such as:

1. Environment data, for instance:
- time stamps and spatial information (when and where was this input received, etc.)
2. Processing information, such as:
- which module has produced this representation; what is its level of confidence, etc.
3. Interactional information:

6. Technical Backgound: XML

At this stage, we should say a word about what appear to be the unavoidable technical choices for the definition of a multimodal content representation format that would be used, among other possibilities, to exchange information between processing modules within a man-machine dialogue system. As a matter of fact, XML, as defined by the World Wide Web Consortium, appears to be the best candidate so far (and probably for quite a long time) to represent information structures intended to be transmitted across a network. In the following section, we give a very brief overview of XML, which we will then use to illustrate some of the principles mentioned above by means of a concrete example.

XML (eXtended Markup Language) is a simplified (but also in some respects enhanced) version of SGML. It provides a syntax for document markup as well as for the description of the set of tags to be used in classes of documents (a so-called DTD, Document Type Definition). An XML document is made of three parts:

- An XML declaration, which, beyond identifying that the current document is an XML one, allows one to declare the character encoding scheme used in the document (e.g. iso-8859-1, utf-8, etc.);
- A document type declaration, which can point to a DTD. This section can be omitted;
- An XML instance corresponding to the actual data represented by the document.

XML makes an important distinction between a *well-formed* document, which only contains the XML declaration and a syntactically conformant instance, and a *valid* one, where the instance is also checked against the associated DTD.

Among other characteristics, we mention the following important properties of XML:

- XML is both Unicode and ISO 10646 compatible[1]
- XML comes along with a specific mechanism, called *namespaces*, allowing one to combine, within the same document, markup taken from multiple sources. This very powerful mechanism, which is in particular the basis for XSLT and XML schemas, allows more modularity in the definition of an XML structure and also to reuse components defined in another context;
- XML provides a general attribute 'xml:lang' to indicate the language used in a given element (see above).

The W3C also provides three very important recommendations for traversing XML documents, namely:

- XPath, which describes a syntax and associated mechanisms to move within a document instance;
- XPointer, which allows one to indicate a location within a document and is based upon the XPath recommendation;
- XLink, which allows one to combine and qualify a set of pointers to describe a link between them.

[1] The W3C has put pressure on both ISO and the Unicode consortium to make sure that they would not diverge in their parallel work on the definition of a universal character encoding scheme.

These three recommendations are important for instance when one wants to relate some information produced by a given processing level and the information that has been used as input for those processes.

Still, it should be noticed that the existence of such a widely recognized *metalanguage* as XML does not solve our problems for representing multimodal content. First, XML by itself does not come with a formal semantics for its tags, and thus does not satisfy the requirement of semantic adequancy. Second, the requirements of flexibility and extensibility forbid us to try to standardize once and for all a precise XML format, but rather think of providing concepts and tools for anyone to be able to design his or her own format, while preserving interoperability conditions with someone else's choices. This is the spirit in which work has already been done within TC37/SC4 for the definition of TMF (Terminological Markup Framework; ISO 16642, under DIS ballot) and which has recently been taken over to deal with morphosyntactic and syntactic annotation (see (Ide & Romary, 2001a and Ide & Romary, 2001b, respectively). The basic assumption that we make is that there exists an entire class of document formats that can be modelled by combining a *metamodel,* that is an abstract structure shared by all documents of a given type (e.g. syntactic annotation document), with a choice of the data categories that may be associated with the various levels of the metamodel. Such a description can be seen as a specification of the document format, which can be instantiated by providing XML representations for the metamodel and the data categories. In such a view, if a community of researchers and implementers agrees on the definition of a reduced set of metamodels for language resources, the actual choice of data categories is left to the responsibility of a specific application. In this framework, the interoperability between formats is ensured by providing a data category registry which gathers, together with precise reference and definition, the various data categories needed for a particular field.

In the case of multimodal content representation we thus advocate that, beyond agreement on the basic components and mechanisms for instance as described in this paper, which could go into the definition of an actual metamodel for content representation, one should not try to standardize a particular XML format more precisely (though we need to make specific choices to illustrate our approach with concrete examples, see below).

7. A simple example

In the following, we illustrate the possible combination of basic components, general mechanisms, and contextual data categories into a multimodal meaning representation. This representation exemplifies the general methodology that we suggested here, by taking up a sample semantic representation derived from an initial example expressed in the ULF+ format (ULF+ is a slightly updated version of a semantic representation language that was developed successively in the PLUS dialogue project, see Geurts and Rentier, 1993, and in the multimodal DENK project; see Bunt et al., 1998; Kievit, 1998).

In the XML excerpt below (corresponding to the sentence "I want to go from Paris to Stuttgart" uttered by a speaker named Peter), we have extended the original ULF+ representation to introduce the notion of dialogue act, whose participants are the speaker and the system. This example is intended to show how we can differentiate between three types of information in such a representation:

- The instantiation of the semantic content representation metamodel as an XML outline (shown in underlined characters), which organizes the general information layout of the data to be represented;
- The actual information units describing the various levels in the XML outline (shown in gray characters);
- The generic mechanisms used to combine events, participants, restrictions and relations (indicated in **bold characters**).

The specific choices made in this example to represent the metamodel or the data categories as XML objects are only one possibility among many, and this does not affect the formal semantics of the underlying information structure. More precisely, the following explanation may help to clarify the example:

- The <semRep> element corresponds to the semantic representation of one elementary utterance or dialogue act. It is identified uniquely by an id attribute;
- The <event> element is used in this example to represent both the dialogue act proper ("e1") and the event expressed by the corresponding linguistic content ("e2");
- The <participant> element is used to represent the various entities involved in the events. Events and participants being related to one another by means of <relation> elements (with source and target attributes pointing to the corresponding arguments of the relation.

The various levels are then further described by a number of data categories, chosen here to illustrate the wide variety of possible cases. Notice the use of an <alt> structure to illustrate the case where an ambiguity would remain at a given step of analysis, each possibility being associated with a certainty evaluation ('cert' attribute). In accordance with the methodology developed in TMF, the name of the corresponding XML elements and attributes should not be the object of standardization, data categories being defined by abstract properties.

```
<semRep id="rep1">
    <event id="e0">
      <evtCat>utterance</evtCat>
      <speaker target="Peter"/>
      <adressee target="System"/>
```

```
    <alt>
        <dialAct cert="0.8">Order</dialAct>
        <dialAct cert="0.3">Inform</dialAct>
    </alt>
</event>

<participant id="Peter">
    <!-- A description of the speaker that can be referendum elsewhere in the document -->
</participant>

<event id="e1">
    <tense>present</tense>
    <voice>active</voice>
    <wh>none</wh>
    <evtType>wanttogo</evtType>
    …
</event>

<participant id="x">
    <lex>I</lex>
    <synCat>Pronoun</synCat>
    <num>sing</num>
    <pers>first</num>
    …
</participant>

<participant id="y">
    <lex>Nancy</lex>
    <synCat>ProperNoun</synCat>
    <pers>third</num>
    …
</participant>

<participant id="z">
    <lex>Stuttgart</lex>
    <synCat>ProperNoun</synCat>
    <pers>third</num>
    …
</participant>

<relation source="x" target="e1">
    <role>agent</role>
</relation>

<relation source="y" target="e1">
    <role>source</role>
</relation>

<relation source="y" target="e1">
    <role>goal</role>
</relation>
</semRep>
```

8. Action Plan

The variety of existing theoretical approaches, as well as the wide number of factors to be considered makes it very difficult to devise from scratch a truly generic framework for multimodal content representation. As a consequence it is necessary to involve, beyond the possibilities offered by the definition of a working group on this topic in TC37/SC4, as large a community of experts as possible in the development of such a framework. This is why we suggest that the work shall be initially conducted within a dedicated working group of SIGSEM (Special Interest Group on Computational Semantics of the Association of Computational Linguistics), which would be, right from the beginning, a liaison with TC37/SC4. This group would prepare a working draft, which would then be submitted to ISO.

Doing so, it would also be easier to ensure a proper interaction with other interested communities, in particular the people working on multimedia representation (SIGMedia, in complement to the existing liaison between MPEG and TC37/SC4) and on discourse and dialogue (SIGDial).

The agenda would thus be the following:

- Refining the workplan on the basis of the present paper at the TC37/SC4 Preliminary Meeting in Jeju (Korea) in February 2002.
- Presenting a position paper at the LREC workshop on "International Standards of Terminology and Language Resources Management" in May 2002.
- First working group meeting in conjunction to IWCS-5 (5th International Workshop on Computational Semantics) in Tilburg, the Netherlands, in January 2003.

9. References

Bunt, H.C., 2000. Dialogue pragmatics and context specification. In H. Bunt and W. Black, editors, *Abduction, Belief and Context in Dialogue.* John Benjamins Publishing Company, Amsterdam.

Bunt, H.C., R. Ahn, R.J. Beun, T. Borghuis and C. van Overveld, 1998. Multimodal cooperation with the DENK system. In: H.C. Bunt, R.J. Beun and T. Borghuis, editors *Multimodal Human-Computer Communication*. Springer, Berlin.

Bunt, H.C., M. Kipp, M.T. Maybury and W. Walster, forthcoming. *Fusion and Coordination for Multimodal Interaction. Roadmap, Arcitecture, Tools, Semantics.*

Geurts, B. and G. Rentier, 1993. Quasi-logical form in PLUS. Internal Report, Esprit Project P5254, *A Pragmatics-based Language Understanding System*. Tilburg University.

Ide, N., A. Kilgariff and L. Romary 2000, A Formal Model of Dictionary Structure and Content. *Proceedings Euralex 2000,* Stuttgart.

Ide, N. and L. Romary, 2001a, Standards for Language Resources, *IRCS Workshop on Linguistic Databases*, 11-13 December 2001, University of Pennsylvania, Philadelphia, USA.

Ide, N. and L. Romary, 2001b, A Common Framework for Syntactic Annotation, *Proceedings of ACL'2001*, Toulouse. Morgan Kaufman, Menlo Park.

Kievit, L.A., 1998. *Context-driven Natural Language Interpretation*. Ph.D. Thesis, Tilburg University.

Maybury, M.T. editor, 1993. *Intelligent Multimedia Interfaces*. AAAI/MIT Press. 405 pp. ISBN 0-262-63150-4 (www.aaai.org:80/Press/Books/Maybury1, mitpress.mit.edu/book-home.tcl?isbn=0262631504)

Maybury, M.T. and W. Wahlster, editors, 1998. *Readings in Intelligent User Interfaces*. Morgan Kaufmann Press. (www.mkp.com/books_catalog/catalog.asp?ISBN=1-55860-444-8)

Romary, L., 2002. MMIL requirements specification. Project MIAMM – *Multidimensional Information Access using Multiple Modalities.* EU project IST-20000-29487, Deliverable D6.1. LORIA, Nancy.

Where will the Standards for Intelligent Computer-Assisted Language Learning Come from?

Lars Borin

Computational Linguistics, Department of Linguistics,
Stockholm University, SE-106 91 Stockholm, Sweden
and
Department of Linguistics, Uppsala University,
Box 527, SE-751 20 Uppsala, Sweden

lars.borin@ling.su.se, lars.borin@ling.uu.se

Abstract

Intelligent computer-assisted language learning—Intelligent CALL, or ICALL—can be defined in a number of ways, but one understanding of the term is that of CALL incorporating language technology (LT) for e.g. analyzing language learners' language production, in order to provide the learners with more flexible—indeed, more 'intelligent'—feedback and guidance in their language learning process. However, CALL, ICALL and LT have been three largely unrelated research areas, at least until recently. In the world of education, 'e-learning' and 'ICT-based learning' are the new buzzwords. Generally, what is meant is some kind of web-based setup, where course materials are delivered via the Internet or/and learners are collaborating using computer-mediated communication (CMC). An important trend in ICT-based learning is that of standardization for reusability. Standard formats for all aspects of so-called 'instructional management systems' are rapidly gaining acceptance in the e-learning industry. Thus, learning applications will need to support them in order to be commercially viable. This in turn means that the proposed standards should be general enough to support all conceivable kinds of educational content and learning systems. In this paper, we will discuss how ICALL applications can be related to the various standards proposals, basing our discussion on concrete experiences from a number of (I)CALL projects, where these standards are used or where their use has been contemplated.

1. Introduction

For some years, I have been actively involved in trying to combine computer-assisted language learning (CALL) with language technology (LT) (a.k.a. computational linguistics (CL), language engineering (LE), or natural language processing (NLP)) into what is often referred to as "Intelligent CALL" (ICALL), both as a teacher of CALL to LT students at the university, and as a researcher involved in a number of research efforts dealing with CALL/ICALL (see below), and also with neighboring areas, such as computer support for lesser used and lesser taught languages (Borin, 2000a; Allwood and Borin, 2001; Nilsson and Borin, 2002), and contrastive linguistic studies using computational methods (Borin, 1999; Borin, 2000b; Borin and Prütz, 2001; Borin and Prütz, 2002).

The present paper flows from a desire to make ICALL benefit from, as well as inform, ongoing standardization efforts in the computational linguistics and e-learning communities.

The rest of the paper is organized in the following way. First, I will try to sort out the relationships between CALL, LT, artificial intelligence (AI), and ICALL. Then I will describe briefly ongoing standardization work in the e-learning and CL communities, and some of the standards proposals that this work has produced. Following that, I will turn to a description of some (I)CALL projects in which I have been or am currently involved, where these standards are used or where their use has been contemplated, namely the SweLL Didax project, the LingoNet project, 'Corpus based language technology for computer-assisted learning of Nordic languages', the SVANTE learner corpus project, and 'IT-based collaborative learning in Grammar'. Finally, I will discuss the situation of ICALL with regard to this standardization work, in order to form an understanding of where we stand at the moment, but more importantly, of where we would like to go from here.

2. CALL, LT and ICALL

Intelligent computer-assisted language learning—Intelligent CALL, or ICALL—has been defined in a number of ways, but one understanding of the term relevant here is that of CALL incorporating LT techniques for e.g. analyzing language learners' language production or modeling their knowledge of a second/foreign language in order to provide them with more flexible—indeed, more 'intelligent'—feedback and guidance in their language learning process.

CALL, ICALL and LT have been three largely unrelated research areas, at least until recently:

1. The CALL 'killer apps' have been e-mail, chat and multimedia programs, developed and used by language teaching professionals with very little input from LT research (Pennington, 1996; Chapelle, 1997; Chapelle, 1999; Chapelle, 2001; Levy, 1997; Salaberry, 1999). The only kind of LT which has had any kind of impact on the CALL field is corpus linguistics, and even in this case it has been the Humanities Computing 'low-tech' kind of corpus linguistics,

rather than the kind pursued in LT (the latter is sometimes referred to as "empirical natural language processing").

2. ICALL has often been placed by its practitioners in the field of artificial intelligence (AI), rather than in LT (e.g. Swartz and Yazdani (1992); Holland et al. (1995)), more specifically in the subfield of AI known as *intelligent tutoring systems* (ITS) (e.g. Frasson et al. (1996); Goettl et al. (1998)). Partly for this reason, work on ICALL has proceeded, by and large, without feedback into the LT community.

3. But on the other hand, in LT in general, (human) language learning has not been seen as an application area worth pursuing. In the recent broad *State of the art of human language technology* overview edited by Cole et al. (1996), 'language learning' does not appear even once in the index, and there is no section on CALL. Certainly there are some exceptions to this general trend; there have been occasional COLING (*International Conference on Computational Linguistics*) papers on ICALL, although few and far between (e.g. Borissova (1988); Zock (1996); Schneider and McCoy (1998)), and there is a research group in Groningen which has been working very actively on LT-based CALL applications for quite some time (Nerbonne and Smit, 1996; Dokter, 1997; Dokter, 1998; Dokter and Nerbonne, 1997; Dokter et al., 1997; Jager et al., 1998). The situation has been changing somewhat only in the last few years, however, with dedicated workshops on language learning applications of CL being arranged in connection with LT conferences and the like (e.g. Olsen (1999); Schulze et al. (1999); Efthimiou (2000)).

3. Standardization in e-Learning and Language Technology

3.1. E-learning standardization efforts

In the world of education, 'e-learning' and 'ICT-based learning'[1] are the new buzzwords (see, e.g., European Commission (2000)). Generally, what is meant is some kind of web-based setup, where course materials are delivered via the Internet or/and learners are collaborating using computer-mediated communication (CMC) methods.

An important trend in ICT-based learning is that of standardization for reusability. Standard formats are defined for all aspects of so-called 'instructional management systems'. Thus, not only educational content formats are agreed upon, but also course structure formats, test formats, as well as how their interaction with recordkeeping systems used in education should take place. There is a number of organizations working on standards in the e-learning area, the most important ones being IMS (Instructional Management System Inc. `http://www.imsproject.org/`), IEEE's LTSC (Learning Technology Standards Committee; `http://ltsc.ieee.org/`), the American Department of Defence ADL (Advanced Distributed Learning; `http://www.adlnet.org/`) initiative, and the European ARIADNE project. Standards being developed by these and other bodies include educational metadata (Learning Objects Metadata – LOM; Anderson and Wason (2000)), test formats (IMS Question and Test Interoperability – QTI; Smythe and Shepherd (2000)), content packaging formats (IMS Content Packaging; Anderson (2000)), modular courseware (ADL SCORM; Dodds (2001)), and others (see, e.g. the IMS and LTSC websites referred to above). At least some of these standards are rapidly gaining acceptance in the e-learning industry. Thus, learning applications will need to support them in order to be commercially viable. This in turn means that the proposed standards should be general enough to support all conceivable kinds of educational content and learning systems.

The general idea is to create standards which are

> "pedagogically neutral, content-neutral, culturally neutral and platform-neutral"
> (Farance and Tonkel, 1999, 9),

and which support. . .

> "common, interoperable tools used for developing learning systems [. . .]
>
> a rich, searchable library of interoperable, "plug-compatible" learning content [. . .]
>
> common methods for locating, accessing and retrieving learning content"
> (Farance and Tonkel, 1999, 14)

One may certainly entertain doubts as to the general attainability of these goals, but one cannot afford to ignore the huge amount of time and labor invested in pursuit of their fulfillment by the organizations mentioned above and others. This being so, it is of course not unimportant if learning and teaching within a particular field—such as language learning—is adequately covered by the proposed standards or not.

3.2. Standardization in Language Technology/Computational Linguistics

In the LT world, too, standardization efforts are legion, and a recurring theme at the LREC (Language Resources and Evaluation Conference) series of conferences.

There is LT standardization work going on at least in the areas of

- resource storage and exchange: TIPSTER (Grishman et al., 1997), ATLAS (Bird et al., 2000), XCES (Ide et al., 2000);
- resource annotation: XCES (Ide et al., 2000), EAGLES (e.g., tagsets: see Monachini and Calzolari (1996));
- resource metadata: OLAC, ISLE (Wittenburg et al., 2000);
- resource presentation and manipulation, and software integration: THISTLE, GATE (Cunningham, 2001), KABA (Olsson, 2002).

[1] ICT is to be read out "Information and Communication Technologies".

To the best of my knowledge, however, the work within LT on resource markup and annotation has not been informed by language learning applications or by the work done on compiling and investigating so-called learner corpora by applied linguistics researchers (see, e.g., Granger (1998)).

4. (I)CALL Case Studies

In this section, we will look at some CALL research projects, where the issue of combining (I)CALL applications with e-learning standards has arisen in various ways.

4.1. Didax

Didax – the Digital Interactive Diagnostic Administering and Correction System, is a project in the framework of the *Swedish Learning Lab* (SweLL), a research effort funded by the Knut & Alice Wallenberg Foundation as part of the larger *Wallenberg Global Learning Network* endeavor, where a number of centers—or "nodes"—worldwide receive funding for exploring the use of ICT and other new technologies in higher education.

At present, there are three nodes in the WGLN: (1) SweLL, with three participating institutions of higher education, (1a) the Royal Institute of Technology and (1b) Karolinska Institutet in Stockholm, and (1c) Uppsala University, (2) the Stanford Learning Lab (SLL), at Stanford University, California, USA, and (3) Learning Lab Lower Saxony (L3S), at the University of Hannover, Germany. SweLL research is currently organized into a multi-tiered structure, with two top-level 'projects' subdivided into a number of 'experiments´. Each experiment is further subdivided into 'tracks', where each track in turn typically is made up of several research teams cooperating on related research issues. Our work on Didax is thus carried out in the *Digital Resources in the Humanities* (DRHum) track of the *Archives – Portfolios – Environments* (APE) experiment of the SweLL project *New meeting places for learning – New learning environments*.

The Didax research team currently consists of three computational linguists and one SLA researcher, but we also cooperate closely with the other DRHum research teams, drawing on the other kinds of competence found there, especially the teams working with digital archives for humanities teaching, as well as with the Uppsala Learning Lab e-folio project group.

The end result of the Didax project is supposed to be a web-based language testing environment, which will provide both students and teachers with a more flexible format for taking, marking, constructing and setting diagnostic language tests in higher education. In Figure 1, the overall architecture of Didax is shown. The three Didax clients (*teacher – setting test*, *teacher – marking test*, and *student*) run in ordinary web browsers. There is nothing out of the ordinary to be seen in any of the client interfaces. This is quite deliberate. Most of the innovation is hidden under the surface, and the interface is a familiar one from many web applications. Didax is described in more detail by Borin et al. (2001).

4.2. LingoNet

LingoNet is a one-year R&D project funded by the Swedish Agency for Distance Education. The project is a cooperation between the Divison of IT Services and the Department of Humanities, Mid Sweden University, and the Department of Linguistics, Uppsala University (see `http://www.mitt.mh.se/lingonet/`).

The aim of the LingoNet project is to build a 'language lab on the Internet', i.e. a web site with a collection of language training resources to be used in higher education, both locally and in distance education. Even though the point of departure for the LingoNet project is the traditional language lab, we actually envision a more general language training resource than this, i.e. a 'computer language lab', rather than a 'computerized version of the tape recorder-based language lab', as the idea is not only to transfer older techniques into this new technology, but also to exploit the additional possibilities offered by the new technology itself, including the incorporation of LT-based language learning resources in the LingoNet lab.

Specifically, in the LingoNet project, we make systematic use of quality control and metadata. It is a well-known fact that the information to be found on the web on any topic is, not only abundant in almost all cases, but also—to put it mildly—of extremely varying quality. At the same time, web search engines are still fairly primitive, so that finding educational resources, appropriate as to their content and level—regardless of their quality—in itself takes some work (Howard Chen, 1999, 24f.). It is only after they have been found that the real work begins, however, when the chaff—resources which are of low quality or of the wrong kind—is to be separated from the wheat—the resources which we can use for our educational purpose, i.e. educational web resources which are quality controlled and classified as to their content and level. In the LingoNet project, the quality control and metadata markup are done by academic language teachers. For more details about the LingoNet project, see Borin and Gustavsson (2000).

4.3. Corpus based language technology for computer-assisted learning of Nordic languages

'Corpus based language technology for computer-assisted learning of Nordic languages', or in short, the Squirrel project, is funded by the Nordic Council of Ministers, and represents a collaboration between the University of Helsinki in Finland, the research foundation SINTEF in Norway, and Stockholm University in Sweden (see `http://www.informatics.sintef.no/projects/CbLTCallNordicLang/squirrel.html`).

One of the aims of the Squirrel project has been to build a prototype web browser for students and teachers of Nordic languages as a second language, which will help them to find practice texts on the web according to the three parameters *language*, *topic*, and *text difficulty* (Nilsson and Borin, 2002). For more details about the Squirrel project, see Borin et al. (2002)

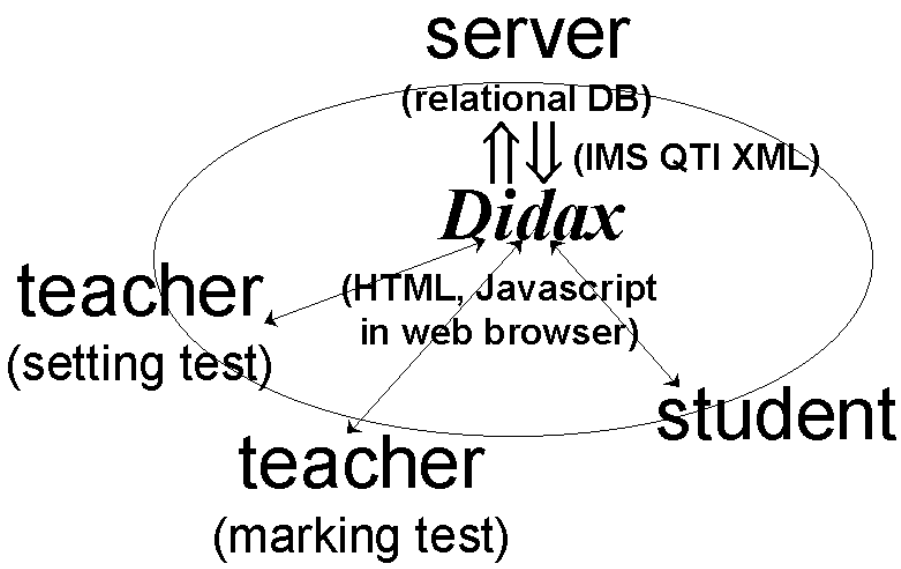


Figure 1: The anatomy of Didax

4.4. SVANTE

SVANTE (SVenska ANdraspråksTexter – Swedish Second Language Texts) is a loose collaboration between linguists, computational linguists, and teachers of Swedish as a second language, with the aim of creating a versatile learner corpus of written Swedish, to complement the learner corpora of spoken Swedish that already exist (see `http://www.ling.uu.se/lars/SVANTE/`). The SVANTE project is partly funded by VINNOVA within the CrossCheck second language Swedish grammar checking project (see `http://www.nada.kth.se/theory/projects/xcheck/`).

4.5. IT-based collaborative learning in Grammar

'IT-based collaborative learning in Grammar' is a collaborative project, funded by the Swedish Agency for Distance Education, with partners in the Linguistics Departments at the universities in Uppsala and Stockholm, and the IT Department and two language departments at Uppsala University. This project revolves around two fundamental assumptions:

1. The use of web-based communication and collaboration technologies will help us make make basic grammar courses better and more effective for students and teachers alike;

2. Language resources originally developed in a research setting, such as tagged and parsed corpora (of Swedish in our case) and grammar writing workbenches, can be (re)used in the context of teaching grammar (Borin and Dahllöf, 1999).

Perhaps I should clarify at this point that this is not primarily an application intended for *language* students, but rather for students of Linguistics and Computational Linguistics, although we believe that it will be useful also as a component in language courses (Saxena and Borin, 2002).

4.6. Relation to e-learning standards and to ICALL

These projects are variously related to ICALL on the one hand and to e-learning standards on the other:

- Didax is not an ICALL project *per se*, but creates an infrastructure which can be used for ICALL applications, and thus must be able to accomodate them. It uses the IMS QTI, and the IEEE, IMS, ARIADNE LOM emerging standards.

- LingoNet is not an ICALL project either, but it goes without saying that among the more exciting possibilities for a web-based language lab are language training applications built on LT methods and resources; hence, we must take this into consideration in designing the underlying language lab format. Like Didax, LingoNet can be considered as an infrastructure project which should be able to accomodate ICALL applications. The standards involved are IMS Content Packaging, and IEEE, IMS, ARIADNE LOM.

- Squirrel is an ICALL project, which does not (yet) utilize any of the proposed e-learning standards, but we see how e.g. the LOM could be used to mark up the located text resources, e.g. for inclusion in something like the LingoNet database.

- SVANTE forms an integral part of an ICALL project, namely the CrossCheck second language grammar checking project, but SVANTE itself is more in the way of a linguistic resource project, where LT standards for basic markup and linguistic annotation of the texts are important.

- 'IT-based collaborative learning in Grammar' is very much an ICALL project. At this initial stage of the project (it started in January 2002), there are still a number of implementational details left to be decided.

However, we would certainly like to make our learning resources as widely useful as possible, meaning, i.a.,

1. that they should be—wholly or in part—easy to integrate into other e-learning environments, but also
2. that it should be easy to use corpus resources for other languages than Swedish in our application.

The first requirement implies the existence and use of general standards for e-learning applications, while the fulfillment of the second requirement certainly would be facilitated by standardization of language resources.

5. So, where will the Standards for ICALL Come from?

Summing up the foregoing, we may say that there are three communities which would benefit from closer interaction, because of a considerable overlap in their goals, but which thus far have pursued these goals separately:

1. The 'ordinary' CALL community—including those researchers working with learner corpora—has extremely tenuous links to LT (see e.g. Chapelle (2001, 32ff.)), and, as far as I have been able to acertain, none at all to the ongoing e-learning standardization work mentioned in section 3.1. above.

2. Nor is the e-learning community working on any standardization for *language learning* (as opposed to *learning* in general). For example, the IMS Question and Test Interoperability (QTI) proposal specifies five test question response types, which can be rendered in up to three different formats (Smythe and Shepherd, 2000, 17). However, for the 'IT-based collaborative learning in Grammar' application, as well as for many other of the corpus-based CALL applications found in the literature, a response type "select (portion/s of) a text" would certainly be good to have.[2]

3. The LT community is not involved in any standardization effort for *language learning* information (as opposed to *language* information in general). The kinds of standards that come to mind first are those involving linguistic annotation schemes, with regard to both their content and their form:

   So-called *learner interlanguage* is characterized by a number of linguistic features absent from the native-speaker version of the target language (and sometimes absent from the learner's native language as well (Richards and Sampson, 1974, 6)). Interlanguage goes through a number of stages, terminating in a final (hopefully close) approximation of the target language. This has some implications for linguistic annotations of learner language production, whether in learner corpora (longer texts) or in analyzers of free learner language production in ICALL language exercises. Thus, part-of-speech (POS) tagging or parsing of learners' interlanguage may have to deal with categories absent from the canonical target language grammar as reflected in an LT standard, etc., but which can be related either to categories in the learner's native language, to universally unmarked categories, to a conflation of target categories, to the pedagogy used, to some combination of these, etc. (Cook, 1993, 18f.). The status of a given linguistic element can change from one language learning stage to another, e.g. the unmarked form in a morphological paradigm becoming functionally more and more specified, as the learner acquires the marked forms and their functions.[3]

   Hence, multiple linguistic annotations of the kind proposed for XCES (Ide et al., 2000) and ATLAS (Bird et al., 2000; Cotton and Bird, 2002) are a necessity for language learning applications of e.g. language corpora.[4] In addition to providing multiple annotations of the same linguistic object (a word, phrase, etc.), the annotations should also be relatable to each other, making it possible to relate an analysis of a form in learner production to the (inferred) intended interpretation of this form, for providing appropriate feedback to the learner. The linguistic categories provided by annotation standards would need to be different from the ones used by native speaker experts (which is arguably most often the kind of annotation aimed for now) if they are to be used for formulating feedback to language learners. They would also have to be different for different kinds of learners, depending on their level, background, native language, etc.

   Standardization of (formats for) *error typologies* would also be desirable. Again, this desideratum is not exclusive to language learning applications; work on grammar and style checkers for native speakers would also benefit from standardized formats for error typologies.

   In the same way as the learner's language progresses through successively more advanced stages, the authentic language that the learner is exposed to as part of her learning process should be successively more complex, in a linguistic sense. This is the main motivation for the Squirrel web search application described above (Nilsson and Borin, 2002). Here, there is consequently a need for a classification and concomitant annotation scheme which relates linguistic complexity to language learning stages, for applications where corpora are used for e.g. generating lan-

[2] In the QTI specification, there is actually a sixth response type response-extension, intended for proprietary response types, but the predefined types will always determine the 'path of least resistance', at least for many users.

[3] Here I have in mind cases such as when e.g. learners of English initially use the infinitive (or sometimes gerund) as their only—and hence extremely polyfunctional—verb form, and then gradually start using other forms (tensed forms in finite clauses, etc.), which then usurp, as it were, some of the functions of the initial forms.

[4] Multiple annotations actually seem necessary for other reasons as well, see e.g. Sampson (2000).

guage learning exercises.

In language learning applications, the need to cater for *bilingual* and *multilingual* text materials is evident, which raises the issues of how to handle multiple writing systems in a standardized way, e.g. left-to-right and right-to-left writing in the same text corpus (the latter issue is raised by Cotton and Bird (2002) as still not having been determined for ATLAS).

Hopefully, the state of affairs depicted here is really due more to lack of interaction than anything else, and if the present paper can be instrumental in bringing about this interaction, it will have served its purpose.

6. Acknowledgements

The work reported herein was carried out partly within the project 'Corpus based language technology for computer-assisted learning of Nordic languages', in the framework of the Nordic Language Technology Research Program 2000–2004 (Holmboe, 2002), funded by the Nordic Council of Ministers through Nordisk Forskeruddannelsesakademi (NorFA), partly within the project 'Digital resources in the humanities', funded by the Knut & Alice Wallenberg Foundation, as part of the Wallenberg Global Learning Network, and partly within the CrossCheck/SVANTE project, funded by VINNOVA within the Language Technology Program.

7. References

Jens Allwood and Lars Borin. 2001. Datorer och språkteknologi som hjälpmedel i bevarandet av romani • Computers and language technology as an aid in the preservation of Romani. Plenary presentation at the symposium *Romani as a language of education: possibilities and restrictions today*. Göteborg University.

Thor Anderson and Tom Wason. 2000. IMS learning resource meta-data information model. final specification version 1.1. Retrieved from the WWW in August 2000: `http://www.imsproject.org/metadata/mdinfov1p1.html`.

Thor Anderson. 2000. IMS content packaging information model. final specification version 1.0. Retrieved from the WWW in October 2000: `http://www.imsproject.org/content/packaging/cpinfo10.html`.

Steven Bird, David Day, John Garofolo, John Henderson, Christophe Laprun, and Mark Liberman. 2000. ATLAS: a flexible and extensible architecture for linguistic annotation. In *Proceedings of LREC 2000*, pages 1699–1706, Athens. ELRA.

Lars Borin and Mats Dahllöf. 1999. A corpus-based grammar tutor for Education in Language and Speech Technology. In *EACL'99. Computer and Internet Supported Education in Language and Speech Technology. Proceedings of a Workshop Sponsored by ELSNET and The Association for Computational Linguistics*, pages 36–43, Bergen. University of Bergen.

Lars Borin and Sara Gustavsson. 2000. Separating the chaff from the wheat: Creating evaluation standards for web-based language training resources. In Khaldoun Zreik, editor, *Learning's W.W.W. Web Based Learning, Wireless Based Learning, Web Mining. Proceedings of CAPS'3*, pages 127–138, Paris. Europia.

Lars Borin and Klas Prütz. 2001. Through a glass darkly: Part of speech distribution in original and translated text. In Walter Daelemans, Khalil Sima'an, Jorn Veenstra, and Jakub Zavrel, editors, *Computational Linguistics in the Netherlands 2000*, pages 30–44. Rodopi, Amsterdam.

Lars Borin and Klas Prütz. 2002. New wine in old skins? A corpus investigation of L1 syntactic transfer in learner language. To be presented at the *International Conference on Teaching and Language Corpora (TaLC) 2002*, Bertinoro, Italy.

Lars Borin, Karine Åkerman Sarkisian, and Camilla Bengtsson. 2001. A stitch in time: Enhancing university language education with web-based diagnostic testing. In *20th World Conference on Open Learning and Distance Education The Future of Learning – Learning for the Future: Shaping the Transition. Düsseldorf, Germany, 01–05 April 2001. Proceedings*, Oslo. ICDE. (CD-ROM: ISBN 3-934093-01-9).

Lars Borin, Lauri Carlson, and Diana Santos. 2002. Corpus based language technology for computer-assisted learning of Nordic languages: Squirrel. Progress report September 2001. In Henrik Holmboe, editor, *Nordisk sprogteknologi. Nordic Language Technology*. Museum Tusculanums Forlag, Københavns Universitet, Copenhagen.

Lars Borin. 1999. Alignment and tagging. In *Working papers in Computational Linguistics & Language Engineering 20*, pages 1–10. Department of Linguistics, Uppsala University.

Lars Borin. 2000a. A corpus of written Finnish Romani texts. In Donncha Ó Cróinin, editor, *LREC 2000. Second International Conference on Language Resources and Evaluation. Workshop Proceedings. Developing Language Resources for Minority Languages: Reusability and Strategic Priorities*, pages 75–82, Athens. ELRA.

Lars Borin. 2000b. You'll take the high road and I'll take the low road: Using a third language to improve bilingual word alignment. In *Proceedings of the 18th International Conference on Computational Linguistics*, pages 97–103, Saarbrücken. Universität des Saarlandes.

Elena Borissova. 1988. Two-component teaching system that understands and corrects mistakes. In *COLING Budapest. Proceedings of the 12th International Conference on Computational Linguistics. Vol I*, pages 68–70, Budapest. John von Neumann Society for Computing Sciences.

Carol Chapelle. 1997. CALL in the year 2000: Still in search of research paradigms? *Language Learning & Technology*, 1(1):19–43. `http://llt.msu.edu/`.

Carol Chapelle. 1999. Research questions for a CALL research agenda: a reply to Rafael Salaberry. *Language Learning & Technology*, 3(1):108–113. `http://llt.msu.edu/`.

Carol Chapelle. 2001. *Computer Applications in Second Language Acquisition*. Cambridge University Press,

Cambridge.

Ron Cole, Joseph Mariani, Hans Uszkoreit, Annie Zaenen, and Victor Zue, editors. 1996. *Survey of the State of the Art in Human Language Technology*. Cambridge University Press, Cambridge. Also as `http://cslu.cse.ogi.edu/HLTsurvey/`.

Vivian Cook. 1993. *Linguistics and Second Language Acquisition*. Macmillan, London.

Scott Cotton and Steven Bird. 2002. An integrated framework for treebanks and multilayer annotations. In *Proceedings of LREC 2002*, Las Palmas. ELRA. To appear.

Hamish Cunningham. 2001. *Software architecture for language engineering*. Ph.D. thesis, University of Sheffield.

Philip Dodds. 2001. ADL SCORM – Advanced Distributed Learning Sharable Content Object Reference Model. Retrieved from the WWW in February 2001: `http://www.adlnet.org/`.

D.A. Dokter and J. Nerbonne. 1997. A session with Glosser-RuG. Alfa-Informatica, University of Groningen. Retrieved from the WWW in November 1998: `http://odur.let.rug.nl/~glosser/welcome.html`.

D.A. Dokter, J. Nerbonne, L. Schurcks-Grozeva, and P. Smit. 1997. Glosser-RuG; a user study. Alfa-Informatica, University of Groningen. Retrieved from the WWW in November 1998: `http://odur.let.rug.nl/~glosser/welcome.html`.

D.A. Dokter. 1997. Glosser-RuG; Prototype December 1996. Alfa-Informatica, University of Groningen. Retrieved from the WWW in November 1998: `http://odur.let.rug.nl/~glosser/welcome.html`.

D.A. Dokter. 1998. From Glosser-RuG to Glosser-WeB. Alfa-Informatica, University of Groningen. Retrieved from the WWW in November 1998: `http://odur.let.rug.nl/~glosser/welcome.html`.

Eleni Efthimiou, editor. 2000. *LREC 2000. Second International Conference on Language Resources and Evaluation. Workshop Proceedings: Language Resources and Tools for Educational Applications*, Athens. ILSP.

European Commission. 2000. e-Learning – designing tomorrow's education. Commission of the European Communities, Communication from the Commission. COM(2000) 318 final. Brussels, 24.5.2000.

Frank Farance and Joshua Tonkel. 1999. LTSA specification. Learning Technology Systems Architecture, draft 5. Retrieved from the WWW in March 2000: `http://edutool.com/architecture/`.

Claude Frasson, Gilles Gautier, and Alan Lesgold, editors. 1996. *Intelligent Tutoring Systems. Third International Conference, ITS '96. Montréal, Canada, June 12–14, 1996. Proceedings*. Number 1086 in Lecture notes in computer science. Springer, Berlin.

Barry P. Goettl, Henry M. Halff, Carol L. Redfield, and Valerie J. Shute, editors. 1998. *Intelligent Tutoring Systems. 4th International Conference, ITS '98. San Antonio, Texas, USA, August 16–19, 1998. Proceedings*. Number 1452 in Lecture notes in computer science. Springer, Berlin.

Sylviane Granger, editor. 1998. *Learner English on Computer*. Longman, London.

Ralph Grishman, Ted Dunning, Jamie Callan, Bill Caid, Jim Cowie, Louise Guthrie, Jerry Hobbs, Paul Jacobs, Matt Mettler, Bill Ogden, Bev Schwartz, Ira Sider, and Ralph Weischedel. 1997. TIPSTER text phase II architecture design. Version 2.3.

V. Melissa Holland, Jonathan D. Kaplan, and Michelle R. Sams, editors. 1995. *Intelligent Language Tutors: Theory Shaping Technology*. Erlbaum, Mahwah, New Jersey.

Henrik Holmboe, editor. 2002. *Nordisk sprogteknologi. Nordic Language Technology*. Museum Tusculanums Forlag, Københavns Universitet, Copenhagen.

Hao-Jan Howard Chen. 1999. Creating a virtual language lab: an EFL experience at National Taiwan Ocean University. *ReCALL*, 11(2):20–30.

Nancy Ide, Patrice Bonhomme, and Laurent Romary. 2000. XCES: an XML-based encoding standard for linguistic corpora. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC2000)*, pages 825–830, Athens. ELRA.

Sake Jager, John A. Nerbonne, and A.J. van Essen, editors. 1998. *Language Teaching and Language Technology*. Swets & Zeitlinger, Lisse.

Michael Levy, editor. 1997. *Computer-Assisted Language Learning. Context and Conceptualization*. Clarendon Press, Oxford.

Monica Monachini and Nicoletta Calzolari. 1996. Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora. a common proposal and applications to European languages. EAGLES Document EAG-CLWG-MORPHOSYN/R.

John Nerbonne and Petra Smit. 1996. GLOSSER-RuG: In support of reading. In *COLING–96. The 16th international conference on computational linguistics. Proceedings, vol. 2*, pages 830–835, Copenhagen. Center for Sprogteknologi.

Kristina Nilsson and Lars Borin. 2002. Living off the land: The Web as a source of practice texts for learners of less prevalent languages. In *Proceedings of LREC 2002*, Las Palmas, Canary Islands, Spain. ELRA. To appear.

Mari Broman Olsen, editor. 1999. *Computer Mediated Language Assessment and Evaluation in Natural Language Processing. A joint ACL–IALL symposium*. Retrieved from the WWW in July 1999: `http://umiacs.umd.edu/~molsen/acl-iall/accepted.html`.

Fredrik Olsson. 2002. *Requirements and Design Considerations for an Open and General Architecture for Information Refinement*. Number 35 in Reports from Uppsala University, Department of Linguistics, RUUL. Uppsala University, Department of Linguistics.

Martha C. Pennington, editor. 1996. *The Power of CALL*. Athelstan, Houston, Texas.

Jack C. Richards and Gloria P. Sampson. 1974. The study of learner English. In Jack C. Richards, editor, *Error Analysis. Perspectives on Second Language Acquisition*. Longman, London.

Rafael Salaberry. 1999. Call in the year 2000: Still developing the research agenda. *Language Learning & Technology*, 3(1):104–107. http://llt.msu.edu/.

Geoffrey Sampson. 2000. Where should annotation stop? In Anne Abeille, Torsten Brants, and Hans Uszkoreit, editors, *Proceedings of the Workshop on Linguistically Interpreted Corpora. LINC-2000*, pages 29–34. Held at the Centre Universitaire, Luxembourg, August 6, 2000.

Anju Saxena and Lars Borin. 2002. Locating and reusing sundry NLP flotsam in an e-learning application. In *Proceedings of LREC 2002 workshop on Customizing Knowledge in NLP Applications: Strategies, Issues, and Evaluation*. To appear.

David Schneider and Kathleen F. McCoy. 1998. Recognizing syntactic errors in the writing of second language learners. In *COLING-ACL '98. Proceedings of the Conference, Vol. II*, pages 1198–1204, Montréal. Université de Montréal.

Mathias Schulze, Marie-Josée Hamel, and June Thompson, editors. 1999. *Language Processing in CALL*. EUROCALL/CTI Centre for Modern Languages, Hull.

Colin Smythe and Eric Shepherd. 2000. IMS question & test interoperability information model specification. version 1.01 – final specification. Retrieved from the WWW in December 2000: http://www.imsproject.org/question/qtinfo101.html.

Merryanna L. Swartz and Masoud Yazdani, editors. 1992. *Intelligent Tutoring Systems for Foreign Language Learning*. Springer, Berlin.

P. Wittenburg, D. Broeder, and B. Sloman. 2000. Metadescription for language resources. EAGLES/ISLE. a proposal for a meta description standard for language resources. Retrieved from the WWW in May 2001: http://www.mpi.nl/world/ISLE/.

Michael Zock. 1996. Computational linguistics and its use in real world: the case of computer assisted-language [sic] learning. In *COLING–96. The 16th International Conference on Computational Linguistics. Proceedings, vol. 2*, pages 1002–1004, Copenhagen. Center for Sprogteknologi.

Personal Names in Unrestricted Chinese Texts: Nature and Identification

Benjamin K. TSOU, Lawrence Y. L. Cheung

Language Information Sciences Research Centre
City University of Hong Kong
Tat Chee Avenue, Kowloon Tong,
Hong Kong

{rlbtsou, rlylc}@cityu.edu.hk

Abstract

The detection of personal names as well as proper names, and the identification of unknown words in unrestricted texts are critical tasks in NLP for East Asian languages, especially for word segmentation, information retrieval and machine translation. This is even more critical for Chinese which uses almost exclusively only the Chinese script and has little overt morphological markings and no equivalent use of capital letters for proper nouns as in English. This paper: (1) discusses the extent of the problems in some relevant IT applications, (2) analyzes the structure of Chinese personal names, and (3) presents some relevant processing strategies and the supporting language resources in general. Differences among Chinese personal names in Beijing and in Hong Kong are highlighted. It is argued that the awareness of variation in names across different Chinese communities constitutes a critical factor in enhancing the effectiveness of Chinese personal name identification algorithms.

Keywords: Chinese, personal name identification, word segmentation, Chinese IT applications, Chinese linguistic differences

1. Introduction

Personal names constitute an important linguistic symbol in conveying meaning. They are anchors of ideas, events, cultural artifacts, etc., e.g. Nobel Prize, Newtonian physics, Clinton-like behaviour and Thatcherism. Personal names provide a rich source for terminology in many domains. Efficiency in personal name identification is important for improving the detection and extraction of terms in the field of computational terminology. The last few years have seen the growth of research in this area (Miller et al., 1999; Cucchiarelli and Velardi, 2001). Named entity recognition was highlighted as an evaluation task in the Sixth and Seventh Message Understanding Conferences (MUC-6 and MUC-7) and First and Second Multilingual Entity Task (MET-1 and MET-2).

Because of the diverse and important linguistic differences between Chinese and English, personal name identification in Chinese involves many more complex issues than in English, e.g., word segmentation, absence of capital/small letter distinction, morphological paucity, syntactic ambiguity, and significant social and cultural differences among Chinese communities (Tsou and Kwong, 2001). Recent statistics from Chinese corpora provides an indicative range of personal names appearing in different domains (Tsou, 2000; Tsou, 2001). Table 1 shows that personal names account for as much as 16.8% of all word types in the 3-year LIVAC[1] newspaper corpus. They represent up to 2.4% of the word tokens in the 29 million character corpus.

	Newspaper Headlines[2]		Newspaper Texts[2]	Court Proceedings
%	Hong Kong (1 yr)	Taiwan (1 yr)	6 Chinese Communities (3 yrs)	Hong Kong (1 case)
Type	4.5	4.2	12.8 to 16.8	4.6
Token	4.4	3.7	1.6 to 2.4	0.6

Table 1 Amount of personal names in different domains

Because of the inherent linguistic problems above, the processing of Chinese personal names (as well as other named entities) in NLP requires much more than itemized listing, and poses a serious challenge.

The rest of the paper will be divided into three main sections: (1) assesses some relevant basic problems encountered in IT applications, (2) introduces the structure of Chinese personal names and the relevant processing strategies, and (3) highlights the importance of building language resources for personal name extraction.

[1] LIVAC synchronous corpus collects newspaper texts every four days since 1995 from Chinese newspapers in 6 Chinese communities including Beijing, Hong Kong, Macau, Shanghai, Singapore and Taiwan.
(LIVAC website: http://www.rcl.cityu.edu.hk/livac)

[2] The estimation is based on the 3 year data (1995—98) from the LIVAC corpus. It contains 29 million characters.

2. Significance of Personal Names in IT Applications

Efficient identification of personal names is crucial in many IT applications. Poor management of personal names in these systems can compound the errors in other NLP modules, resulting in serious deterioration of system performance. Cheung et al. (2002) conducted tests showing that poor personal name processing results in serious webpage retrieval errors.

(1) 陳中將與俄羅斯選手爭奪跆拳道金牌 (Sina) [3]
Chen Zhongjiang will compete with a Russian athlete for the gold medal for Taekwondo.

(2) 司令陳中將視導南測中心受訓部隊 (Google Chi.)[3]
Admiral Chen Zhongjiang inspected the trainee army force in Nance centre.

For example, the examined search engines mistook 中將 *zhongjiang* in (1) and (2) as the common noun for the military rank of "lieutenant general", whereas, in fact, they represent given names in the above contexts. The problem is similar to identifying *Dean Martin*, the well known American entertainer, as the head of a faculty in a university.

Tsou and Kwong (2001) also reported that the Chinese-to-English machine translation systems [4] have serious but unrecognized problems handling personal names. Table 2 shows that all four machine translation systems perform rather poorly in personal name identification. The probable cause for the errors is the use of static name list to identify personal names. The above demonstrates that IT applications need far more sophisticated algorithms than simple character matching and name database to adequately detect personal names in Chinese texts.

Data Source	Translation Accuracy			
	EWGate	*TongYi*	*Transtar*	*WorldLingo*
Hong Kong	24%	5%	9%	15%
Beijing	30%	6%	20%	56%
Taiwan	19%	0%	5%	16%

Table 2 Translation accuracy of personal names

3. Processing Chinese Personal Names: Challenges and Strategy

3.1. Challenges in Processing Chinese Personal Names

The basic structure of modern Chinese personal names is largely similar across different Chinese communities. Although the frequent length is 2 to 3 characters, the maximum can be as long as 6 characters. Table 3 shows the possible structures of Chinese personal name. Chinese personal names begin with a one- or two-character surname, followed by a one- or two-character given name. The name of a married female may be preceded by her husband's surname, as in (e) and (f). The unique structure

	Full Name	Husband's Surname			Surname			Given Name		Length
		H1	(H2)	+	S1	(S2)	+	G1	(G2)	
a.	李鵬 *Li Peng*				李 *Li*			鵬 *Peng*		2
b.	鄧小平 *Deng Xiaoping*				鄧 *Deng*			小 *Xiao*	平 *Ping*	3
c.	諸葛亮 *Zhuge Liang*				諸 *Zhu*	葛 *Ge*		亮 *Liang*		3
d.	東方聞櫻 *Dongfang Wenying*				東 *Dong*	方 *Fang*		聞 *Wen*	櫻 *Ying*	4
e.	陳方安生 *Chen Fang Ansheng*	陳 *Chen*			方 *Fang*			安 *An*	生 *Sheng*	4
f.	諸葛東方聞櫻 *Zhuge Dungfang Wenying*	諸 *Zhu*	葛 *Ge*		東 *Dong*	方 *Fang*		聞 *Wen*	櫻 *Ying*	6

Table 3 Structure of Chinese personal names

[3] *Google* [Big5 Chinese] URL: http://www.google.com/intl/zh-TW and *Sina* URL: http://www.sina.com.cn

[4] (1) Transtar V3.0, (2) TongYi '98, (3) *WorldLingo* (http://www.worldlingo.com), (4) EWGate: (http://www.EWGate.com/ewtranslite.html)

is found in speech or writing of formal register in some Chinese communities such as Hong Kong.

Apart from variable length, several characteristics make Chinese personal name processing difficult:

(a) There is no explicit morphological marking or capitalization for names in Chinese.

(b) Chinese texts do not have explicit word boundary.

(c) The character set for surnames and given names is a subset of Chinese characters for common Chinese words, and hence readily gives rise to structural ambiguity.

(d) Some personal names may be simple mono-syllabic words.

(e) Some polysyllabic words can be embedded in Chinese personal names, e.g. 王朝聞 *Wang Chaowen* (王朝 *wangchao* = dynasty), 馬勝利 *Ma Shengli* (勝利 *shengli* = victory) and 嚴肅 *Yan Su* (嚴肅 *yansu* = serious(ly)).

3.2. Basic Strategies

The complexity of Chinese personal name identification task calls for a combination of different processing strategies. They can be broadly divided into statistical approach and linguistic approach.

3.2.1. Linguistic Approach

Linguistic context provides important cues to locate Chinese personal names. Syntactic structures and lexical collocation provide good indication on whether or not the character string immediately before or after it is a potential personal name, e.g. 張志偉先生 (*Mr.* Zhang Zhiwei) and 朱鎔基總理 (*Premier* Zhu Rongji). Sun et al. (1995) integrates features to detect frequently used patterns, lexical items and syntactic structures that are useful for identifying names. For example, personal names often precede verbs like 說 *shuo* (say), 指出 *zhichu* (point out), etc. Lü et al. (2001) detect personal names by evaluating the interaction between potential personal names and neighbouring words. The POS co-occurrence restriction is checked and the best segmentation for potential name string is computed so as to generate the most probable context. Luo and Song (2001) studied the structure of personal name and place name formation. The linguistic knowledge is represented as a set of generative rules in finite state automata. Additional exceptional handling is added to deal with easily confused ambiguous contexts.

3.2.2. Statistical Approach

Statistical approach has been the most popular approach for name identification task. Previous studies typically exploited the character distribution frequency in different parts of a name and designed algorithms to extract string patterns that match the distributional criteria. For example, Sun et al. (1995) and Song and Tsou (2001) reported that about 400 characters[5] could cover over 99% of all Chinese surnames in texts. Furthermore, some character combinations in given names are more frequent than others. Cheung et al. (2002) also pointed out that there are significant variations among Chinese communities. The character preference in given names varies depending on a range of factors like gender, geography, character position in a given name, social changes, etc. The character probability is approximated by frequency distribution from large text corpora or name databases.

Sun et al. (1995) and Zheng et al. (1999) evaluated every candidate string by computing mutually exclusive probability for the 3 characters in a name candidate string, as in (6).

(6) $p_{pn}(c_1\, c_2\, c_3) = p_{sur}(c_1) * p_{m1}(c_2) * p_{m2}(c_3)$

where
$p_{pn}(s)$ = probability of candidate string s being a personal name
$p_{sur}(x)$ = probability of character x being a surname
$p_{m1}(x)$ = probability of character x being the first character of a given name
$p_{m2}(x)$ = probability of character x being the second character of a given name

Lü et al. (2001) proposed to measure probability of a potential name string by considering the probability of the 3 characters in a name candidate string as mutually inclusive events, as in (7) adapted from Lü et al. (2001).

(7) $p_{pn}(c_1\, c_2\, c_3) = p_{1F}(c_1) + p_{1M}(c_2) + p_{nE}(c_3)$

where
$p_{pn}(s)$ = probability of candidate string s being a personal name
$p_{1F}(x)$ = probability of character x being a surname
$p_{1M}(x)$ = probability of character x being the first character of a given name
$p_{nE}(x)$ = probability of character x being the second character of a given name

Most Chinese personal name identification algorithms incorporate linguistic and statistical techniques. These hybrid systems have been reported to achieve 80—90% precision and recall rates (Sun et al., 1995; Lü et al., 2001; Luo and Song, 2001).

[5] There are 21,886 characters in the GBK Chinese character set.

4. Personal Name Language Resources for Terminology Extraction

Statistical frequency data, as discussed in Section 3, has to be based on empirical data from large text corpora. Thus relevant personal name databases become a critical resource to support name identification systems and to customize algorithms. At least four major dimensions should be adequately addressed in the construction of personal name language resources, including: (1) structural distribution, (2) character frequency of personal names, (3) character co-occurrence for given names, and (4) communal differences. The significance and relevance of appropriate personal name database cannot be overemphasized because of the rarely understood magnitude of variation of personal names among Chinese communities which is much greater than that existing in English speaking communities. We will illustrate the differences in personal name patterns by using name databases taken from Beijing and Hong Kong.[6]

4.1. Structural Distribution

Single-character surnames predominate both databases, accounting for over 99%, as in Table 4. This suggests that double-character surnames may be handled separately using item listing in view of its very limited number of types and tokens. The data shows a divergence in the preference for single- and double-character given names in Beijing and in Hong Kong. Single-character names account for 29% of the Beijing database.[7] In contrast, single-character given names only cover 2% of the data for Hong Kong. The findings are crucial to the prioritization of rules related to the length of personal names in identification algorithms.

	Surname		Given Name	
%	Beijing	HK	Beijing	HK
Single-Character	99.9	99.6	29.1	2.1
Double-Character	0.1	0.4	70.9	97.9

Table 4 Distribution of name structures

[6] The Beijing name database has 125,033 names, and is drawn from a county in Beijing. They are representative of names in Mainland China because the county population is composed of migrants coming from different provinces of China. The Hong Kong database contains 11,358 names. They are student and staff names taken from the Registrar's Office, City University of Hong Kong.

[7] Sun et al. (1995) reported that single-character given names account for about 37% of the name database for all students' names (10 years) at Tsinghua University in Beijing.

4.2. Character Frequency of Personal Names

Not all characters are equally probable in being different parts of a Chinese personal name. All studies mentioned in Section 3 have exploited such characteristics to different extent. Table 5, 6 and 7 show that the ten most frequently used surnames, first character and second character of given names.

	Beijing				Hong Kong		
Rank	Surname	%	Cum. %	Rank	Surname	%	Cum. %
1	王 *Wang*	9.1	9.1	1	陳 *Chen*	10.2	10.2
2	張 *Zhang*	8.3	17.4	2	黃 *Wang*	6.7	16.9
3	李 *Li*	7.9	25.3	3	李 *Li*	5.9	22.8
4	劉 *Liu*	6.5	31.8	4	梁 *Liang*	4.6	27.4
5	陳 *Chen*	3.2	35.0	5	林 *Lin*	4.2	31.6
6	趙 *Zhao*	3.2	38.2	6	張 *Zhang*	3.6	35.2
7	楊 *Yang*	3.0	41.2	7	劉 *Liu*	3.0	38.2
8	孫 *Sun*	2.0	43.2	8	吳 *Wu*	3.0	41.2
9	馬 *Ma*	1.7	44.9	9	何 *He*	2.8	44.0
10	吳 *Wu*	1.6	46.5	10	鄭 *Zheng*	2.1	46.1

Table 5 10 most frequent single-character surnames in Beijing and Hong Kong

(Shaded items appear in both columns.)

	Beijing				Hong Kong		
Rank	G1	%	Cum. %	Rank	G1	%	Cum. %
1	淑 *shu*	3.2	3.2	1	嘉 *jia*	3.8	3.8
2	玉 *yu*	3.1	6.3	2	偉 *wei*	3.7	7.5
3	秀 *xiu*	2.9	9.1	3	志 *zhi*	3.5	11.0
4	曉 *xiao*	2.6	11.7	4	家 *jia*	2.8	13.8
5	文 *wen*	2.3	14.0	5	詠 *yong*	2.2	16.0
6	建 *jian*	2.2	16.2	6	慧 *hui*	2.1	18.1
7	志 *zhi*	1.9	18.0	7	國 *guo*	2.0	20.1
8	小 *xiao*	1.8	19.8	8	文 *wen*	2.0	22.0
9	桂 *gui*	1.7	21.5	9	佩 *pei*	1.9	24.0
10	春 *chun*	1.4	22.8	10	麗 *li*	1.9	25.9

Table 6 10 most frequent first characters (G1) of double-character given names

(Shaded items appear in both columns.)

	Beijing				Hong Kong		
Rank	G2	%	Cum. %	Rank	G2	%	Cum. %
1	華 *hua*	3.6	3.6	1	儀 *yi*	3.1	3.1
2	英 *ying*	3.4	7.0	2	華 *hua*	2.3	5.4
3	蘭 *lan*	2.1	9.1	3	明 *ming*	2.2	7.6
4	平 *ping*	1.9	11.0	4	敏 *min*	2.2	9.8
5	珍 *zhen*	1.8	12.8	5	文 *wen*	2.1	11.9
6	明 *ming*	1.7	14.5	6	玲 *ling*	1.9	13.7
7	榮 *rong*	1.6	16.1	7	珊 *shan*	1.7	15.4
8	生 *sheng*	1.5	17.6	8	欣 *xin*	1.6	17.0
9	芳 *fang*	1.3	18.9	9	輝 *hui*	1.6	18.6
10	琴 *qin*	1.3	20.1	10	雯 *wen*	1.6	20.1

Table 7 10 most frequent second characters (G2) of double-character given names

(Shaded items appear in both columns.)

The character type for Chinese surnames is fairly limited in actual data. In Table 5, the ten most frequent surnames cover over 46% of the name tokens though the ranking of surnames is quite different in both databases. For example, the most frequent surname 王 *Wang* in Beijing is ranked as 14th in the Hong Kong. In contrast, the character types for given names are far more diverse. In the Hong Kong database, there are over 820 character types for given names as opposed to 257 character types for surnames. As shown in Table 6 and 7, the ten most frequently used G1 and G2 character cover no more than 26% of all name tokens in both databases respectively.

4.3. Character Co-occurrence for Given Names

Apart from localized character preference in given names, our data also reveals that the character combinations of double-character given names are far from being random. Previous research tended to consider the probabilities of each character position in isolation, and ignored interesting patterns of character co-occurrence in given names. The information is useful for resolving ambiguity given rise by the diverse character types in given names. Here are two examples for the two most common G1 characters from Hong Kong database: 嘉 *jia* and 偉 *wei*. Given G1 = 嘉 *jia* / 偉 *wei*, there is about 30% of chance that the given name is one of the combinations in a—e and f—j respectively. (Table 8)

	Combina-tion	%	Cum. %		Combina-tion	%	Cum. %
a	嘉 + 敏 *jia + min*	11.2	11.2	f	偉 + 強 *wei + qiang*	6.4	6.4
b	嘉 + 儀 *jia + yi*	6.5	17.7	g	偉 + 文 *wei + wen*	5.7	12.1
c	嘉 + 雯 *jia + min*	4.6	22.3	h	偉 + 雄 *wei + xiong*	5.7	17.7
d	嘉 + 琦 *jia + qi*	4.3	26.6	i	偉 + 傑 *wei + jie*	*5.4*	23.2
e	嘉 + 慧 *jia + wei*	3.6	30.1	j	偉 + 明 *wei + ming*	5.2	28.3

Table 8 *5* most frequent combinations provided G1 = 嘉 *jia* / 偉 *wei*

4.4. Communal Differences

Previous studies do not seem to pay much attention to the sociolinguistic aspects of name variation among Chinese communities. It has been mentioned in Section 4.1 that there are far more single-character given names in the Beijing database. If we further compare the columns for Beijing and Hong Kong in Table 6 and 7, only two characters overlap. The divergence in character preference is obvious in the two databases. The implication is that name identification algorithms using statistical approach should maintain character probability derived from various Chinese communities in order to maximize the performance. Other sociolinguistic differences such as married female's names, nicknames, etc, have yet to be studied. The assumption that personal name identification can be simplistically tackled on the basis of personal name language resources from a single community will certainly be problematical for NLP applications that have to process unrestricted texts from different geographical locations.

5. Further Works

Based on Section 4.2 and 4.3, further investigation into the statistical distribution of personal names can be done. First, it seems that previous studies tended to have overlooked character co-occurrence phenomenon. Co-occurrence probability can be used to improve existing algorithms for Chinese personal name tagger. For example, instead of merely utilizing the probability of a candidate character being part of a name, the tagger may give a higher rating to those candidate strings whose G1 and G2 combination is commonly found in Chinese names. Statistical studies like those in Section 4.3 will be conducted for all character combination in the two databases to identify high frequency patterns.

Second, as we mentioned earlier and noted by a reviewer, gender is a significant factor in character choice for given names. Such data may find applications in transcription system and speech recognition application such as caller name identification. The recognition engine may first determine the caller's gender based on the speaker's voice pitch and then select the appropriate probability database for name identification task accordingly.

Third, the communal differences revealed by our preliminary analysis suggest that name databases from other Chinese communities are important language resources. For example, more name databases will be collected (e.g. Shanghai, Taiwan and Singapore) for comparison.

6. Conclusion

Personal names provide an important source for new terms in text processing. Personal name identification is crucial to terminology extraction. This paper discusses the challenge and basic strategies in personal name identification in unrestricted Chinese texts. The review of IT applications shows that reliability in the processing of Chinese personal names is still far from acceptable. This situation contributes to serious errors in other NLP tasks such as incorrect data retrieval and parsing. Current Chinese personal name identification systems capitalize on linguistic and statistical techniques to deal with the processing. To adequately support such systems, personal name language resources are critical. Four dimensions have been highlighted in the construction of such resources, including (1) structural distribution, (2) character frequency of personal names, (3) character co-occurrence for given names, and (4) communal differences. Despite the potential contribution to the identification task, the latter two dimensions seem to have gone largely unnoticed in the literature. More empirical study of personal names will be beneficial to the performance improvement of personal name identification systems.

7. Acknowledgement

This research study is supported by the Language Information Sciences Research Centre, City University of Hong Kong and by a Competitive Earmarked Research Grant (CityU 1238/00H) from the Research Grant Council of Hong Kong and supported by NTT Service Integration Laboratory. We would also like to thank Rou SONG for his kindness in providing us with the Beijing personal name database, and Registrar's Office, City University of Hong Kong, for their contribution to our Hong Kong personal name database.

8. References

Cheung, L., B. K. Tsou and M. Sun. (2002) Identification of Chinese Personal Names in Unrestricted Texts. *Proceedings of the 16th Pacific Asia Conference on Language, Information and Computation*, Cheju, Korea, pp. 28—35.

Cucchiarelli, A. and P. Velardi. (2001) Unsupervised Named Entity Recognition Using Syntactic and Semantic Contextual Evidence. *Computational Linguistics* 27 (1): 123—131.

Luo, Z. and R. Song. (2001) Integrated and Fast Recognition of Proper Noun in Modern Chinese Word Segmentation. [in Chinese] *Proceedings of International Conference on Chinese Computing 2001*, Singapore. pp. 323—328.

Lü, Y., T. Zhao, M. Yang, H. Yu and S. Li. (2001) Leveled Unknown Chinese Words Resolution by Dynamic Programming. [in Chinese] *Journal of Chinese Information Processing*, 15 (1), Beijing, China.

Miller, D., R. Schwartz, R. Weischedel, and R. Stone. (1999) Named entity extraction from broadcast news." *Proceedings of DARPA Broadcast News Workshop*. Herndon, VA.

Song, R. and B. K. Tsou. (2001) Preliminary Study on Chinese Proper Noun. [in Chinese] *Proceedings of the 20th Anniversary Conference of the Chinese Information Processing Society of China*. November 2001. pp. 14—19.

Sun M., C. Huang, H. Gao and J. Fang. (1995) Identifying Chinese Names in Unrestricted Texts. [in Chinese] *Journal of Chinese Information Processing*, 9 (2), Beijing, China.

Tsou, B. K. (2000) Lexical Variation in Chinese: The Windows Approach. (Invited paper) Annual Research Forum, Linguistic Society of Hong Kong. December 2000.

Tsou, B. K. (2001) Corpus, Information Mining and the New Global Village. (Keynote speech) *Proceedings of 6th Natural Processing Pacific Rim Symposium*, Tokyo, November 2001. pp. 9—18.

Tsou, B.K. and Kwong, O.Y. (2001) Evaluating Chinese-English Translation Systems for Personal Name Coverage. *Proceedings of the MT Summit VIII Workshop on MT 2010 -- Towards a Road Map for MT*, Santiago de Compostela, Spain.

Zheng, J., X. Lin and H. Tan. (2000) The Research Chinese Names Recognition Method Based on Corpus. [in Chinese] *Journal of Chinese Information Processing*, 14 (1), Beijing, China.

Changes in the Etymological Type of New Terminology in Japanese
-The Decrease of Sino-Japanese and Increase of Alphabetical Terms-

SHIODA, Takehiro

NHK Broadcasting Culture Research Institute
Mori Tower 16F, 2-5-1
Tokyo 105-6216, Japan
sioda@culture.nhk.or.jp

Abstract

In this paper, the author surveys how the proportions of etymological type have changed in current computer-related terms of Japanese. As a result of inquiry regarding recent computer terminology, the fact that the decreasing of Sino-Japanese words and the increasing of Alphabetical words has confirmed.

Introduction

Japanese words are conventionally divided into three etymological types, known as *goshu* in Japanese, according to whether they are of pure Japanese, Sino-Japanese or Western-loans. In this paper, the author surveys how the proportions of each type have changed in current computer-related terms, and considers what the future may hold for Japanese technical terms.

First, some technical terms required for the analysis are explained, and previous studies introduced. The materials and analytical method are then described, and the results reported. The results of opinion polls on people's thoughts regarding the use of foreign words are also introduced, and a proposal on word coining is presented.

1 "GOSHU " in Japanese Linguistics

The term *goshu* refers to a basic convention used in classifying the parts of the Japanese vocabulary. It is the taxonomical concept for defining words according to their etymological source. The three basic types that are taken to constitute the Japanese vocabulary are the words of pure Japanese, Sino-Japanese and Western-loan.

The pure Japanese words, *wago*, are the words of traditional Japanese origin. These are frequently found in terms that express fundamental concepts in Japanese. They are written in *hiragana* syllabary or *kanji* (Chinese characters) in general.

The Sino-Japanese words, *kango*, can primarily be described as words that are borrowed from Chinese. However, the *kango* are read in a Japanese, not a Chinese way, despite the use of the Chinese characters. (This is similar to the many different pronunciations of the word *euro*, which varies so much from language to language.)
There is also the concept of *wasei-kango*, namely Sino-Japanese words created in Japan, as a subdivision of *kango*. These are unique Japanese coinages that use Chinese morphemes.
The traditional scientific terms include many Sino-Japanese words. It is usual to use Chinese characters when writing these Sino-Japanese terms. It has been observed that weight of the Sino-Japanese words in Japanese language is similar to that of words of Latin origin in English (Miyajima, 1995).

The Western-loans, called *gairaigo*, are mostly loan words from Western languages (mainly English), and sometimes the words of not-western origin are also included exceptionally. The newest terms include many words of western origin. It is usual to use *katakana* syllabary when writing these Western-loans, but the alphabet is also used in some cases.

These three types compose the fundamental taxonomy of Japanese etymological word types.

In addition, some new words are formed by combining the different types. These hybrid words are called *konshugo*.

2 Previous research

It has been shown in quantitative terms that the use of *kango*, Sino-Japanese words, was chiefly utilized in new coinages around 1900, and that ratio gradually decreased thereafter (Miyajima, 1967). This tendency has continued in recent years and the word-formation capability of *kango* fell sharply in the very short period from 1960 to 1980 (Nomura, 1984). As for writing means, it has been predicted that the use of Chinese characters will decrease and that of the alphabet will increase from now on (Kabashima, 1981).

3 Purpose of inquiry

In order to predict future transitions in Japanese terminology, the present situation was gauged with reference to the following points:

1. It has been observed that the word-formation capability of *kango* has been decreasing. What is the rate of this decrease?

2. It is known that the proportion of Western-loans is increasing in Japanese. The author believes that the increase may be greatest for alphabetical words. Can this be demonstrated quantitatively?

In this paper, the author reports the results obtained regarding computer terminology.

4 Procedure of inquiry

Subject of inquiry:
"Gendai Yoogono Kiso Chisiki (Basic knowledge of contemporary words)" 1985, 1990, 1995, 2000: Tokyo, Jiyuu Kokuminsha.

This book is a single volume encyclopedia published annually. It provides rich data for considering the status and progress of new words from year to year. For this study, the entries related to the computer field (computer terms, office automation terms, etc.) were extracted.
Each entry was classified according to the *goshu* category.

1. We observe the transitions of Sino-Japanese and Western-loans in the first, *goshu* classification phase. Since there are very few pure Japanese words, these are disregarded here. "*Katakana* words" and "alphabetical words" are provided as sub-classifications of Western-loans, and the transitions for each are noted. Here, we only observe the number of entries belonging to single *goshu* categories. *Konshugo* are taken up in the second phase below.

2. Next, consideration is also given to the hybrid words, *konshugo*. (This can be described as classification by *goshu* element). Since there are also very few elements of pure Japanese words here, these are again disregarded.

A hybrid word consisting, for example, of one Sino-Japanese and one Western-loan is counted once in each of the Sino-Japanese element and Western-loan element categories. For convenience, however, a term consisting of multiple Sino-Japanese elements is counted only once in the Sino-Japanese element category.

Examples:
「情報検索」(information retrieval):
Scores 1 for the Sino-Japanese element
「エレクトロニック・バンキング」
(electronic banking):
Scores 1 for the *Katakana* word element
「磁気ディスク装置」 (magnetic disk unit):
Scores 1 for the Sino-Japanese element
Scores 1 for the *Katakana* word element
「ＯＣＲ」(Optical Character Recognition):
Scores 1 for the alphabetical word element
「双方向ＣＡＴＶ」(two-way CATV):
Scores 1 for the Sino-Japanese element
Scores 1 for the alphabetical word element

5 Results and discussion

5.1 1st phase (classification by *goshu*)

	1985	*1990*	*1995*	*2000*
Total	398	344	357	402
Sino-Japanese	54(13.6%)	46(13.4)	51(14.3)	37(9.2)
Katakana	136(34.1)	108(31.4)	114(31.9)	118(29.4)
Alphabetical	19(4.8)	30(8.7)	27(7.6)	66(16.4)

(*see* Figure 1)
→ The rates for Sino-Japanese words and alphabetical words were substantially reversed from 1995 to 2000.

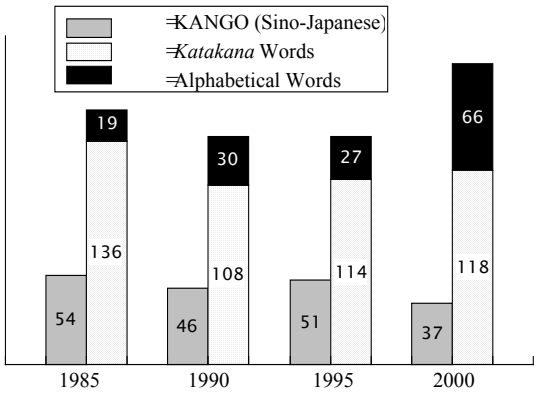


Figure 1: The transition by GOSHU in computer te

5.2 2nd phase (classification by *goshu* element)

	1985	*1990*	*1995*	*2000*
Total	398	344	357	402
Sino-Japanese	218(54.8)	186(54.1)	193(54.1)	155 (38.6)
Katakana	305(76.6)	239(69.5)	251(70.3)	270(67.2)
Alphabetical	66(16.6)	86(25.0)	79(22.1)	158 (39.3)

(*see* Figure 2)

→ The rates for Sino-Japanese and alphabetical elements drew much closer to each other in the data for 1995 to 2000.

Prospect: The likelihood of a further increase in the rate of use of alphabetical words appears to be quite strong.

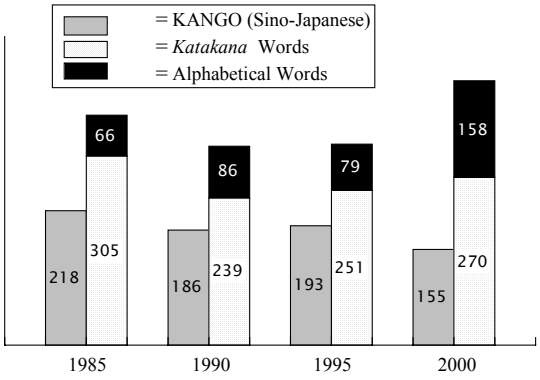


Figure 2: The transition by GOSHU element

6 Views on Western-loans

The excessive use of words of foreign origin can hinder communication. We next introduce some results on this subject from public opinion polls.

"Do you feel that many loan words and other foreign words are used in everyday Japanese?"
Frequent: 51.6%
Occasional: 32.2%
(Agency for Cultural Affairs, 2000)

"Have you been troubled because you cannot understand the meaning of a *katakana* word in newspaper or TV?"
Frequently : 17.1%
Occasionally : 37.5%
(Agency for Cultural Affairs, 1997)

The entry of new foreign terms cannot be prevented. But, as these surveys indicate, we should be aware of the dangers of excess.

7 Concluding remarks

It has been observed that one of the merits of the increase in foreign words is the acceptance of terms that are understood internationally (Ishiwata, 2001). Alphabetical words, in particular, can be read and understood by those who cannot read Japanese script, so the level of international communicability is very high. The risk is that more fluent international communication may be matched by weaker internal communication. The use of such words as technical terms has clear merits, but thought is also required to the selection of words that are best able to acquire general acceptability within the specific language-speaking group concerned. We should remember that *not* all the people understand English.

Some technical terms do gradually come to be used as general terms in each language. Those who coin new terms or standardize the terminology would, therefore, be well advised to consider their suitability for both international and internal communication purposes, with the awareness that these decisions may have some future influence on general terms kept clearly in mind.

References

Agency for Cultural Affairs (Bunkachoo). (1997, 2000). *Kokugoni kansuru yoron choosa (Census on the Japanese Language)*. Tokyo: Ookurashoo insatsukyoku.

Inoue, Fumio. (2001). English as a Language of Science in Japan. From Corpus Planning to Status Planning. *The Dominance of English as a Language of Science -Effects on Other Languages and Language Communities*. Berlin/New York: Mouton de Gruyter.

Ishiwata, Toshio. (2001). *Gairaigo no soogooteki kenkyuu (Comprehensive Study on Western-origin borrowed vocabulary)*. Tokyo: Tookyoodoo shuppan.

Kabashima, Tadao. (1981). *Nihongo wa doo kawaruka (How does Japanese change?)*. Tokyo: Iwanami Shoten.

Miyajima, Tatsuo. (1967). Kindai-goi no keisei (Formation of the modern vocabulary). *Kokuritsu kokugo kenkyuujo ronshuu (Collected Papers of The National Language Research Institute) 3*. Tokyo: Shuuei Shuppan.

Miyajima, Tatsuo. (1995). A Contrastive Study of Vocabulary Growth in Different Languages - French, English, Chinese, and Japanese. *Lexical Knowledge in the Organization of Language*. Amsterdam/ Philadelphia: John Benjamins Publishing Company.

Nomura, Masaaki. (1984). Goshu to zoogoryoku (The etymological type and the capability of word-formation). *Nihongogaku 3-9*, 1984.9. Tokyo: Meiji Shoin.

Shioda, Takehiro. (2000). Japanese and Korean Terminologies Reviewed from a Linguistic Perspective. *Proceedings of Workshop on Terminology Resources and Computation*, Held in conjunction with the LREC2000. Athens.

Shioda, Takehiro. (2002). Senmon-yoogo ni okeru arufabetto-go no zooka (The Increase of Alphabetical Words in Japanese Terminology). *Journal of Japan Society of Information and Knowledge,* Vol.12, No.1. Tokyo: Japan Society of Information and Knowledge.

A Corpus-based Approach to Term Bank Construction

Bai Xiaojing, Hu Junfeng, Zan Hongying, Chen Yuzhong, Yu Shiwen

Institute of Computational Linguistics, Peking University, China

E-mail: {baixj, hujf, zanhy}@pku.edu.cn

Abstract

In this paper, a corpus-base approach is presented in the construction of the information science and technology term bank in which domain classification, reference and part of the definition are extracted from corpus. Farther experiments show that the structure analysis of the terms can be helpful in the corpus-based domain classification of the terms.

1. Introduction

Currently, a joint project is under way between China National Institute of Standardization(CNIS) and the Institute of Computational Linguistics(ICL), Peking University to construct a term bank in the field of information science and technology. The project aims at :

1. an ontology system
2. a corpus for term bank construction
3. a corpus-based terminology extraction program
4. a constructed term bank and the related specifications and standards, and others for terminologies in the field of information science and technology

The implementation of the whole project features various approaches, among which the corpus-based one constitutes our present focus.

The corpus in this project consists of two parts, an essential corpus of 15 million Chinese characters and an extension corpus of 60 million and more, responsible for different tasks respectively. The corpus-based approach enables us to address the goals of our project by the following schemes:

1. Categorization of the terminologies in our term bank
2. Assistance for defining the terminologies in our term bank
3. Training and testing of the automatic extraction program

Now, initial plans have been made for the implementation of these schemes, with experiments conducted in support of our further efforts.

2. The Classification Scheme of Information science and Technology

An ontology system is very important for the standardization of the term bank establishment. Up to now, there still do not have a ready-made classification scheme of information science and technology, not to say to put each specific terminology into one specific domain category. So the first thing for constructing the term bank in the field of information science and technology is to build an appropriate knowledge category system or concept system.

The information science and technology field contains not only the computer and communication subjects. In general, this field includes all subjects relative to information. Now there is no acknowledged opinion that bounds this field. We intend to set up an appropriate and practical classification while make it integrated with the some existed international or national standards. We have referred to the ACM Computing Classification System, ICS(the International Classification for Standards), CLC(the Chinese Library Classification), computer encyclopedias, and some technical dictionaries. After we have consulted many materials, we classify the

knowledge of information science and technology field into five subjects:

1. pandects of information science and technology
2. computer
3. automatization
4. telecommunication
5. electronics

under each subject we provide four subclass: theory, technology, application and product & material. We also have set up a mapping between ICS and our classification. For example, ICS:35:220 are integrate into our classification in data storage device(its classification number is 020403).

Generally, our classification is on the second level of subjects, and some detail on the third or fourth level. Frankly, Our knowledge classification system has fewer hierarchical levels. The reason is that we plan to get a more general and shallow classification and to avoid the frequent modification of the structure of the term bank due to the slight change of term category. The change of terms' intension and extension will be reflected through some attributes in our term bank. The attributes in the term bank are very easily modified or expanded.

3. Corpus Compilation

For the essential corpus, we turn to experts in the field of information science and technology. All the texts are chosen and provided by experts of specified branches.

In the meaning time, with the help of a program, field experts will tag all the terms and the related information in the corpus, i.e., categorize them into the very branches of the field they belong to. The essential corpus is built for data training in the automatic extraction program.

For the extension corpus, the size is more than 60 million Chinese characters. In this corpus, we can get concordance and collocation information about the terms, as automatic processing will be possible for this part, and further, considerable amount of useful information, which can facilitate the definition of the terms, can be extracted from the corpus. Moreover, this corpus will serve as a test set for the terminology extraction program.

4. Corpus-based Categorization of Terminologies

Up till now, a basic framework has been drafted out for the purpose of categorization, while the terminologies available now are more than 70,000. Given the possibility that the initial framework can be developed to a sound system for categorization, locating the Terms into this system will still be a hard job.

It is in this consideration that we come up with the corpus-based approach. The essential corpus provided by various field experts carries the field information and the terminology tagging. Terminologies tagged by field experts are to be compared with the Terms. This is designed to be a process of matching, after which the Terms can be put into their respective categories. In other words, we try to classify the terms according to their distribution in the corpus. For the first step, as a test, we obtained 100 texts (258,045 characters in total) about Computer Network, with 2,486 different terms tagged out (i.e., 2,486 terminologies are regarded as valid). Considerable terms, which are unlikely network ones, proved otherwise in the corpus.

For example:缓冲/cache, which does not seem to be an OS term in Chinese, is a true network concept in the following sentence: "与我们熟悉的磁盘缓冲技术类似，Internet 缓冲是在一台本地服务器上开辟一块缓冲区，保存访问 Internet 时获得的数据，这样在以后的浏览过程中如果还是访问那些网页，就不需要再次访问Internet ，而直接从缓冲中获得数据就可以了".

That means corpus based categorization can give a more accurate description of the field information about the terms. This will benefit not only the term categorization, but also the definition of the terms. In some cases, it can

even give us clues to find out terms with different shades of meaning.

5. Corpus-based Reference for Terminology Definition

Accuracy and standardization in defining terminologies also attract our attention and efforts. In the database of our term bank, there is a field named Reference, storing contexts of the Terms from the whole corpus, which are deemed as competent reference. Reference for terminology definition can be at various levels, namely, it can be sentence(s), paragraph(s) or even full text(s). Here the role of the corpus is significant, as it contains all the information that will be filled into the Reference field, and what is more, we are expecting templates for terminology reference or even for terminology definition, to be learned from the essential corpus and then applied to the extension part, thus achieving the corpus-based automatic referencing. In addition to category and terminology tagging, our field experts also have to tag the text contents that they regard as the competent references for terminologies. A program is designed to extract a language unit bearing a reference tag (starting with <Reference> and ending with </Reference>) containing or following a terminology tag (starting with <Term> and ending with </Term>), which is recognized as the reference information for the tagged terminology and will then be stored in the Reference field accordingly. The following are three examples.

Example 1: (a single sentence)

<Reference><Term>Vo I P **</Term>**可以定义为以IP 包交换的方式传输话音。**</Reference>**

Example 2: (a paragraph)

<Reference><Term>Vo I P 网关**</Term>**主要提供PSTN 电话通信网络与IP 网络的接口和转换。目前，一般采用H.323 作为IP 网络信令和SS7 作为PSTN 的信令。在这个市场的设备提供商中既有传统的数据网络公司如3Com 、Cisco 等，也有老牌的电信设备提供商如Alcatel 、Ericsson 、Nortel 、Lucent 等，以及Sonus 、Clarent 、convergent network 、Nuera 等公司。**</Reference>**

Example 3: (a full text)

<Reference>何谓**<Term>**DHC**P</Term>**？动态主机配置协议（Dynamic Host Configuration Protocol，DHCP）从原有的BootP 协议发展而来，原来的目的是为无盘工作站分配IP 地址的协议，当前更多地用于对多个客户计算机集中分配IP 地址以及IP 地址相关的信息的协议，这样就能将IP 地址和TCP/IP 的设置统一管理起来，而避免不必要的地址冲突的问题，因此常常用在网络中对众多DOS/Windows 计算机的管理方面，节省了网络管理员手工设置和分配地址的麻烦。中继代理服务器必须知道DHCP 服务器的地址，还要知道如何把接收到的报文转发给该服务器**</Reference>**

Sufficient data will avail us of the opportunity to learn reference templates, like “XX 可以定义为/can be defined as XX” in Example1; “XX 主要提供/is mainly for XX” in Example 2 etc. These are sample templates that can be used to extract the definition of the terms from corpus. Surely there can only have small number of the terms that can find definition directly from corpus, but the corpus-based contextual information, such as concordance and collocation are also helpful for experts to analysis the meaning and give the proper definition of the terms.

6. Automatic Extraction of Terminologies from Corpus

The third scheme is based on the understanding that the internal structure of terminologies is also a source of valuable knowledge for term bank construction. In this project, the internal structure of a terminology consists of three elements: 1) term constituents, including prefixes, suffixes, words and phrases that are frequently used in related technical documents, e.g., “性” and “接口”; 2) POS; and 3) semantic categories, each describing the common feature of a group of term constituents, like

the semantic category “equipped with/without a system of wires” derived from “无线” and “有线 ”. Patterning the internal structure of terminologies is a prerequisite to the automatic extraction of terminologies from the corpus. On the one hand, we analyze the Terms, together with those from the essential corpus and tagged by our field experts, and pattern their structures, using term constituents and POS information, e.g., “noun + 接口”. On the other hand, we generate new terms, replacing term constituents of the same categories in exiting terms with the other.

With “有线通讯”, “有线电视”, “有线电报”, for instance, we generate “无线通讯”,“无线电视”, “无线电报”. The automatic extraction program will then use the structure patterns and the new terms generated to extract terminologies from the extension corpus, either by character matching or by POS matching, or both. Large in amount as they are, the terminologies we have obtained r reach up till now. In this sense, the extension corpus is both a test set for the automatic extraction program and a source for additional terminologies by using the program. It therefore are still far from being enough. Considering the limited sources, we have to rely on the extension corpus for the automatic extraction of terminologies that remain out of oucalls our attention to the competence and performance of our corpus, and especially, the extension part.

7. Conclusion

We have devised the initial schemes for the application of the corpus-based approach to

1. the categorization of existing terminologies in our term bank
2. the learning of reference templates and the extraction of reference information from the corpus
3. the modeling of automatic terminology extraction

Experiments show that corpus can be very useful to illuminate the meaning of terms, which will help a lot to standardize the terms in the future.

References

1. Angelo, Robert. *A Synopsis of Wittgenstein’s Logic of Languag*e. http://www.roangelo.net/logwitt.
2. Feng, Zhiwei, , (1997). *An Introduction to Modern Terminolog*y. Yuwen Press
3. Sinclair John, (1991). *Corpus, Concordance, Collocation.* Oxford University Press.
4. Kennedy Graeme, (2000). *An Introduction to Corpus Linguistic*s. Foreign Language Teaching and Research Press
5. http://www.acm.org/class/1998/
6.http://www.iso.ch/iso/en/CatalogueListPage.CatalogueList
7. Chinese Library Classification, Version4.0, Beijing library press, China
8. Zan Hongying, Hu Junfeng, et al (2002) , Construction of the Term Bank, TAHK2002

Standards for Language Resources

Nancy Ide,* Laurent Romary†

* Department of Computer Science
Vassar College
Poughkeepsie, New York 12604-0520 USA
ide@cs.vassar.edu
† Equipe Langue et Dialogue
LORIA/INRIA
Vandoeuvre-lès Nancy, FRANCE
romary@loria.fr

Abstract

This paper presents an abstract data model for linguistic annotations and its implementation using XML, RDF and related standards; and to outline the work of a newly formed committee of the International Standards Organization (ISO), ISO/TC 37/SC 4 Language Resource Management, which will use this work as its starting point. The primary motive for presenting the latter is to solicit the participation of members of the research community to contribute to the work of the committee.

1. Introduction

The goal of this paper is two-fold: to present an abstract data model for linguistic annotations and its implementation using XML, RDF and related standards; and to outline the work of a newly formed committee of the International Standards Organization (ISO), ISO/TC 37/SC 4 Language Resource Management, which will use this work as its starting point. The primary motive for presenting the latter is to solicit the participation of members of the research community to contribute to the work of the committee.

The objective of ISO/TC 37/SC 4 is to prepare international standards and guidelines for effective language resource management in applications in the multilingual information society. To this end, the committee will develop principles and methods for creating, coding, processing and managing language resources, such as written corpora, lexical corpora, speech corpora, dictionary compiling and classification schemes. The focus of the work is on data modeling, markup, data exchange and the evaluation of language resources other than terminologies (which have already been treated in ISO/TC 37). The worldwide use of ISO/TC 37/SC 4 standards should improve information management within industrial, technical and scientific environments, and increase efficiency in computer-supported language communication.

2. Motivation

The standardization of principles and methods for the collection, processing and presentation of language resources requires a distinct type of activity. Basic standards must be produced with wide-ranging applications in view. In the area of language resources, these standards should provide various technical committees of ISO, IEC and other standardizing bodies with the groundwork for building more precise standards for language resource management.

The need for harmonization of representation formats for different kinds of linguistic information is critical, as resources and information are more and more frequently merged, compared, or otherwise utilized in common systems. This is perhaps most obvious for processing multi-modal information, which must support the fusion of multimodal inputs and represent the combined and integrated contributions of different types of input (e.g., a spoken utterance combined with gesture and facial expression), and enable multimodal output (see, for example, Bunt and Romary, 2002). However, language processing applications of any kind require the integration of varieties of linguistic information, which, in today's environment, come from potentially diverse sources. We can therefore expect use and integration of, for example, syntactic, morphological, discourse, etc. information for multiple languages, as well as information structures like domain models and ontologies.

We are aware that standardization is a difficult business, and that many members of the targeted communities are skeptical about imposing any sort of standards at all. There are two major arguments against the idea of standardization for language resources. First, the diversity of theoretical approaches to, in particular, the annotation of various linguistic phenomena suggests that standardization is at least impractical, if not impossible. Second, it is feared that vast amounts of existing data and processing software, which may have taken years of effort and considerable funding to develop, will be rendered obsolete by the acceptance of new standards by the community. To answer both of these concerns, we stress that the efforts of the committee are geared toward defining *abstract models* and *general frameworks* for creation and representation of language resources, rather than specific formats. These models should, in principle, be abstract enough to accommodate diverse theoretical approaches. The model so far developed in ISO TC/37 for terminology, which has informed and been informed by work on representation schemes for dictionaries and other lexical data (Ide, *et al.*, 2000) and syntactic annotation (Ide & Romary, 2001) demonstrates that this is not an unrealizable goal. Also, by situating all of the standards development squarely in the framework of XML and related standards such as RDF, DAML+OIL, etc., we hope to ensure not only that the standards developed by the committee provide for compatibility with established and widely accepted web-based technologies, but also that

transduction from legacy formats into XML formats conformant to the new standards is feasible.

ISO/TC 37/SC 4 will liaison with ISLE (International Standards for Language Engineering), which has implemented various recent efforts to integrate EC and US efforts for language resources. Where possible, these and other standards set up in EAGLES will be incorporated into the ISO standards. ISO/TC 37/SC 4 will also broaden the work of EAGLES/ISLE by including languages (e.g. Asian languages) that are not currently covered by EAGLES/ISLE standards.

At present, language professionals and standardization experts are not sufficiently aware of the standardization efforts being undertaken by ISO/TC 37/SC 4. Promoting awareness of future activities and rising problems, therefore, will be a crucial factor in the success of the committee, and will be required to ensure widespread adoption of the standards it develops. An even more critical factor for the success of the committee's work is to involve, from the outset, as many and as broad a range of potential users of the standards as possible. This presentation serves as a call for participation to the linguistics and computational linguistics research communities.

3. Objectives

ISO TC37/SC 4's goal is to develop a platform for the design and implementation of linguistic resource formats and processes in order to facilitate the exchange of information between language processing modules. This will be accomplished by defining a *common interface format* capable of representing multiple kinds of linguistic information. The interface format must support the communication among all modules in the system, and be adequate for representing not only the end result of interpretation, but also intermediate results.

A well-defined representational framework for linguistic information will also provide for the specification and comparison of existing application-specific representations and the definition of new ones, while ensuring a level of interoperability between them.

3.1. Requirements

Very generally, a linguistic representation framework must meet the following requirements:

- *Expressive adequacy:* the framework should be expressive enough to represent all varieties of linguistic information;
- *Semantic adequacy:* the representation structures should have a formal semantics, i.e., their definition should provide a rigorous basis for further processing (e.g., deductive reasoning, statistical analysis, generation, etc.).

Providing interface formats within a system architecture demands that "incremental" construction of intermediate and partial representations be supported. In addition, if the construction of a final representation does not succeed, the representation must capture the information required to enable appropriate system action. This dictates additional requirements:

- *Incrementality:* support for the various stages of input interpretation and output generation, allowing both early and late fusion and fission.
- *Uniformity:* the representation of various types of input and output should utilize the same "building blocks" and the same methods for combining complex structures composed of these building blocks.
- *Underspecification and partiality*: support for the representation of partial and intermediate results, including the capture of unresolved ambiguities.

Finally, the representational framework must be accommodate the developing field of language processing system design by satisfying these further requirements:

- *Openness:* the framework should not depend on a single linguistic theory, but should enable representations based on different theories and approaches;
- *Extensibilty*. The framework should be compatible with alternative methods for designing representation schemas (e.g., XML) rather than being tied to a specific schema.

3.2. Methodology

A working group of SC 4 (WG1/WI-1) has been charged with the task of defining a linguistic annotation framework, which will be used by other SC 4 working groups to develop more precise specifications for particular annotation types. The full list of SC 4 working groups is as follows:

- WG1/WI-0: Terminology for Language Resources
- WG1/WI-1: Linguistic annotation framework
- WG1/WI-2: Meta-data for multimodal and multilingual information
- WG2/WI-3: Structural content representation (syntax and morphology)
- WG2/WI-4: Multimodal content representation
- WG2/WI-5: Discourse level representation
- WG3/WI-6a: Multilingual text representation
- WG4/WI-7: Lexicons
- WG5/WI-8: Validation of language resources
- WG5/WI-9: Net-based distributed cooperative work for the creation of LRs

We focus here on the work of WG1/WI-1, which will serve as the starting point for that of most of the others. This group will propose a *data architecture* consisting of basic mechanisms and data structures for linguistic annotation and representation, comprised of the following:

- *Basic components*: the basic constructs for building representations of linguistic information; specifically, identification of types of building blocks and ways to connect them.
- *General mechanisms*: representation techniques that make the annotations more compact and flexible and enable linking them to external sources of information; for example, sub-structure labeling, argument under-specification, restrictions on label values and/or disjunctions or lists to represent ambiguity or partiality, structure sharing; linking to

domain models, linking to other levels of annotation, etc.

- *Contextual data categories*: administrative (meta-) data relevant for processing, such as environment data (e.g., time stamps, spatial information); processing information (e.g., module that produced the representation; confidence level); interaction information (speaker, audience, etc.).

The following section outlines a linguistic framework which will serve as the starting point for development within SC 4. The current model is based on work on development of annotation formats for lexicons (Ide, et al., 2001), morphosyntactic and syntactic annotation (Ide & Romary, 2001a; Ide & Romary, 2001b; Ide & Romary, forthcoming), and which has been further developed within TC37/SC4 for the definition of TMF (Terminological Markup Framework; ISO 16642, under DIS ballot).

4. A Framework for Linguistic Annotation

Our fundamental assumption is that representation formats for linguistic data and its annotations can be modeled by combining a structural *meta-model,* that is, an abstract structure shared by all documents of a given type (e.g. syntactic annotation), with a set of *data categories* that are associated with the various components of the meta-model. Our work in SC4 is concerned, first, with identification of a reduced set of meta-models that can be used for any type of linguistic data and its annotations. Data categories, on the other hand, are defined by the implementer; interoperability among formats is ensured by providing a *Data Category Registry* in which the categories and relations required for a particular type of annotation are precisely defined.

The model for linguistic annotation must satisfy two general criteria:

1. It must be possible to instantiate it using a standard representational format;
2. It must be designed so as to serve as a pivot format into and out of which proprietary formats can be transduced, in order to enable comparison and merging, as well as operation on the data by common tools.

4.1. Abstract model for annotation

At its highest level of abstraction, an annotation is a set of data or information (in our case, linguistic information) that is associated with some other data. The latter is what could be called "primary" data (e.g., a part of a text or speech signal, etc.), but this need not be the case; consider, for example, the alignment of parallel translations, where the "annotation" is a link between two primary data objects (the aligned texts). Typically, primary data objects are represented by "locations" in an electronic file, for example, the span of characters comprising a sentence or word, or a point at which a given temporal event begins or ends (as in speech annotation). As such, at the base primary data objects are relatively simple in their structure; more complex data objects may consist of a list or set of contiguous or non-contiguous locations. Annotation objects, on the other hand, often have a more complex internal structure: syntactic annotation, for example, may be expressed as a tree structure, and may include more elemental annotations such as dependency relations (which is itself an annotation relating two objects, where the relation is directional (dependent-to-head)).

Thus, we can conceive of an annotation as a one- or two-way link between an annotation object and a point (or a list/set of points) or span (or a list/set of spans) within a base data set. Links may or may not have a semantics--i.e., a type--associated with them. Points and spans in the base data may themselves be objects, or sets or lists of objects. This abstract formulation can serve as the basis for defining a general model for linguistic annotation that can be realized in a standard representational format. In fact, this model is consistent with well-established data modeling concepts used in diverse areas, including knowledge representation (KR), object-oriented design, and database systems, and which inform fundamental data structures in computer science (trees, graphs, etc.) and database design (notably, the Entity-Relationship (ER) model). As such, the model provides us with established means to describe our data objects (in terms of composition, attributes, class membership, applicable procedures, etc.) and relations among them, independent of their instantiation in any particular form. It also ensures that standardized representation formats exist that can instantiate the model.

One way to represent linguistic annotation in terms of the abstract model is as a graph of elementary *structural nodes* to which one or more *information units* are attached. The distinction between the structure of annotations and the informational units of which it is comprised is, we feel, critical to the design of a truly general model for annotations. Annotations may be structured in several ways; perhaps the most common structure is hierarchical. For example, phrase structure analyses of syntax are structured as trees; in addition, hierarchy is often used to break annotation information into sub-components, as in the case of lexical and terminological information.

There are several special relations *among* annotations that must be represented in the model, including the following:

- *Parallelism:* two or more annotations refer to the same data object;
- *Alternatives:* two or more annotations comprise a set of mutually exclusive alternatives (e.g., two possible part-of-speech assignments, before disambiguation);
- *Aggregation:* two or more annotations comprise a list or set that should be taken as a unit.

Information units or *data categories* provide the semantics of the annotation. Data categories are the most theory and application-specific part of an annotation scheme. We do not attempt to define the relevant data categories for given types of annotation. Rather, we propose the development of a Data Category Registry to provide a framework in which the research community can formally define data categories for reference and use in annotation. To make them maximally interoperable and consistent with existing standards, data categories can be defined using RDF schemas to formalize the properties and relations associated with each. Note that RDF descriptions function much like class definitions in an

object-oriented programming language: they provide, effectively, templates that describe how objects may be instantiated, but do not constitute the objects themselves. Thus, in a document containing an actual annotation, several objects with the same type may be instantiated, each with a different value. The RDF schema ensures that each instantiation is recognized as a sub-class of more general classes and inherits the appropriate properties.

A formally defined set of categories will have several functions: (1) it will provide a precise semantics for annotation categories that can be either used "off the shelf" by annotators or modified to serve specific needs; (2) it will provide a set of reference categories onto which scheme-specific names can be mapped; and (3) it will provide a point of departure for definition of variant or more precise categories. Thus the overall goal of the Data Category Registry is not to impose a specific set of categories, but rather to ensure that the semantics of data categories included in annotations (whether they exist in the Registry or not) are well-defined and understood.

5. An Example

We illustrate a simple application of the framework presented above for the domain of morpho-syntactic annotation. For the purposes of illustration, it is necessary to make technical choices concerning the representation format. XML and related standards developed by the World Wide Web consortium appear at present to provide the best means to represent information structures intended to be transmitted across a network. For the purposes of linguistic resource representation, XML provides several important features:

- it is both Unicode and ISO 10646 compatible;
- XML namespaces provide the options of combining element definitions from multiple sources in an XML document, thereby fostering modularity and reuse;
- XML schemas provide a powerful means to define, constrain, and extend definitions of the structure and contents of classes of XML documents and document sub-parts;
- W3C has defined accompanying standards for inter- and intra-document linkage (XPath, XPointer, and Xlink) as well as document traversal and transformation (XSLT);
- XML is fully integrated with emerging standards such as the Resource Definition Framework (RDF) and DAML+OIL, which can be "layered" on top of XML documents to provide a formal semantics defining XML-instantiated objects and relations.

We have defined an XML format for representing linguistic annotations called the *Generic Mapping Tool (GMT*). The GMT defines XML elements for encoding annotation structure (primarily, a nestable `<struct>` element) and data categories (a nestable `<feat>` tag). A `<seg>` element provides a pointer to the annotated data using XPointers. Relations among objects can be specified explicitly using a `<rel>` element or may be implicit in the hierarchical nesting of `<struct>` elements. The GMT is described in detail in Ide & Romary, 2001b. We stress, however, that the details of the XML format—in particular, element names—is arbitrary; the only requirement is that the underlying data model can be expressed using the format.

5.1. Morpho-syntactic annotation

Morpho-syntactic annotation provides a good example of how the data model instantiated in the GMT is applied, and demonstrates some of the mechanisms required for representing annotations in general. Morpho-syntactic annotation involves the identification of word classes over a continuous stream of word tokens. The annotations may refer to the segmentation of the input stream into word tokens, but may also involve grouping together sequences of tokens or identifying sub-token units (or morphemes), depending on the language under consideration and, in particular, the definitions of "word" and "morpheme" as applied to this language. The description of word classes may include one or several features such as syntactic category, lemma, gender, number etc., which is again dependent on the language being analyzed.

Morpho-syntactic annotation can be represented by a single type of structural node (named W-level) representing a word-level structure unit. One or several information units are associated with each structural node.

For the purposes of illustration, we identify the following data categories (in practice these would be defined in reference to categories in the Data Category Registry):

- /lemma/: contains or points to a reference word form for the token or sequence of tokens being described;
- /part of speech/: a reference to a morpho-syntactic category;
- /confidence/: a confidence level assigned by the manual or automatic annotator in ambiguous cases.
- /gender/: the grammatical gender information associated with a word token or a sequence of word tokens;
- /number/: the grammatical gender information associated with a word token or a sequence of word tokens;
- /tense/: the grammatical tense information associated with a word token or a sequence of word tokens;
- /person/: the grammatical person information associated with a word token or a sequence of word tokens.

The following provides an example of the morpho-syntactic annotation of the sentence "Paul aime les croissants" in the GMT format:[1]

```
<struct type="MSAnnot">
 <struct type="W-level">
   <feat type="lemma">Paul</feat>
   <feat type="pos">PNOUN</feat>
      <seg target="#w1"/>
 </struct>
 <struct type="W-level">
   <feat type="lemma">aimer</feat>
   <feat type="pos">VERB</feat>
   <feat type="tense">present</feat>
   <feat type="person">3</feat>
   <seg target="#w2"/>
 </struct>
```

[1] For brevity, we use an abbreviated pointer syntax to refer to the primary data in this example.

```
<struct type="W-level">
  <feat type="lemma">le</feat>
  <feat type="pos">DET</feat>
  <feat type="number">plural</feat>
  <seg target="#w3"/>
</struct>
<struct type="W-level">
  <feat type="lemma">croissant</feat>
  <feat type="pos">NOUN</feat>
  <feat type="number">plural</feat>
  <seg target="#w4"/>
</struct>
</struct>
```

Note that there is no limit to the number of information units that may be associated with a given structural node (as opposed to the text based representations that are usually provided by available POS taggers). It is also possible to structure the annotations by embedding <feat> elements to reflect a more complex feature-based annotation, or by pointing to a lexical entry providing the information.

In some cases, the morpho-syntactic annotation of a word or sequence of words requires a hierarchy of word level structures (e.g., when a word token results from the combination of several morphemes that must be annotated independently). For example, some occurrences of the token "du" in French can be analyzed as the fusion of the preposition "de" with the determiner "le" (as in "la queue *du* chat"). This is handled by embedding word-level structures as follows:

```
<struct type="W-level">
  <seg target="#w1"/>
  <struct type="W-level">
     <feat type="lemma">de</feat>
     <feat type="pos">PREP</feat>
  </struct>
  <struct type="W-level">
     <feat type="lemma">le</feat>
     <feat type="pos">DET</feat>
  </struct>
</struct>
```

Conversely, annotation of compound words may involve associating a single lemma to a sequence of word tokens at the surface level. In this case, the lemma is attached to the higher level of embedding and reference to the source is given at the leaves of the hierarchy, as in the following representation of the compound "pomme de terre" in French :

```
<struct type="W-level">
  <feat type="lemma">
        pomme_de_terre</feat>
  <feat type="pos">NOUN</feat>
  <struct type="W-level">
     <seg target="#w1"/>
     <feat type="lemma">pomme</feat>
     <feat type="pos">NOUN</feat>
  </struct>
  <struct type="W-level">
     <seg target="#w2"/>
     <feat type="lemma">de</feat>
     <feat type="pos">PREP</feat>
  </struct>
  <struct type="W-level">
     <seg target="#w3"/>
     <feat type="lemma">terre</feat>
     <feat type="pos">NOUN</feat>
  </struct>
</struct>
```

The ability to specify a hierarchical structure where needed enables specification of the level of granularity required. This is especially critical for a representation scheme, since the granularity of the segmentation in (or associated with) the primary data may not directly correspond to the level of granularity required for the annotation.

5.1.1. Alternatives

Morpho-syntactic annotation can be used to illustrate the representation of both structural and informational alternatives, which arises when a given word token is associated with two or more word classes. For example, the French word "bouche" which can be derived both from the verb "boucher" and the noun "bouche", which can be represented as follows:

```
<struct type="W-level">
  <seg target="#w1"/>
  <alt>
    <feat type="lemma">boucher</feat>
    <feat type="pos">VERB</feat>
    <feat type="tense">present</feat>
    <feat type="confidence">0.4</feat>
  </alt>
  <alt>
    <feat type="lemma">bouche</feat>
    <feat type="pos">NOUN</feat>
    <feat type="confidence">0.6</feat>
  </alt>
</struct>
```

5.1.2. Relating annotation levels

We assume the use of stand-off annotation; that is, an annotated corpus is represented as a lattice of stand-off annotation documents pointing to a primary source or intermediate annotation levels. However, depending on the point of view, the relations between various annotation levels can be more or less explicit. It is possible to identify three major ways to relate different levels of annotation: temporal anchoring, event-based anchoring, and object-based anchoring.

Temporal anchoring associates positional information to each structural level. This positional information is typically represented as a pair of numbers expressing the starting point and ending point of the segment being described. To do so in our framework, we introduce two attributes for the <seg> element:

- /startPosition/: the temporal or offset position of the beginning of the current structural node;
- /endPosition/: the temporal or offset position of the end of the current structural node.

For example, the following associates a phonetic transcription with a given portion of a primary text:

```
<struct type="phonetic">
  <seg startsAt="2300"
       endsAt="3200"/>
  <feat type="phone">iy</feat>
</struct>
```

We also define an event-based anchoring, which effectively introduces a structural node to represent a location in the text, to which all annotations for the object

at that location can refer. This strategy is useful in two cases:

- Situations where it is not possible or desirable to modify the primary data by inserting markup to identify specific objects or points in the data (e.g., speech annotation, associated with a speech signal, or in general any "read-only" data).
- Primary data marked with "milestones", such as time stamps in speech data, where spans across the various milestones must be identified. In this case, the `<struct>` elements represent the markup for segmentation (e.g., segmentation into words, sentences, etc.).[2]

To represent this, we introduce a specific type of structural node, named *landmark*, which is referred to by annotations for the defined span, as follows:

```
<struct type="landmark">
    <seg startsAt="2300"
         endsAt="3200"/>
</struct>
```

The third mechanism, object-based anchoring, enables pointing from a given level to one or several structural nodes at another level. This mechanism is particularly useful to make dependencies between two or more annotation levels explicit. For example, syntactic annotation can refer directly to the relevant nodes in a morpho-syntactically annotated corpus, in order, for example, to identify the correct NP "le chat" in "la queue du chat", as shown below:

```
<!-- Morphosyntactic level -->
<struct type="W-level">
   <seg target="#w3">
   <struct type="W-level">
     <seg target="#w3.1">
     <feat type="lemma">de</feat>
    <feat type="pos">PREP</feat>
    </struct>
     <struct type="W-level">
       <seg target="#w3.2">
        <feat type="lemma">le</feat>
       <feat type="pos">DET</feat>
       <feat type="gender">masc</feat>
    </struct>
  </struct>
   <struct type="W-level">
       <seg target="#w4">
       <feat type="lemma>chat</feat>
       <feat type="pos">NOUN</feat>
  </struct>
</struct>
<!-- Syntactic level (simplified) -->
<struct>
   <feat type="synCat">NP</feat>
   <seg targets="w3.2 w4"/>
</struct>
```

[2] The annotation graph (AG) formalism (Bird and Liberman, 2001) was explicitly designed to deal with time-stamped data. However, we feel the AG is not sufficiently general because (1) AG reifies the "arc" and distinguishes it from identification of spans via, e.g., XML tags; and (2) AG requires *ad hoc* mechanisms to deal with hierarchically organized annotations. In both cases, AG requires different mechanisms to treat analogous constructs.

5.2. Summary

The framework presented here for linguistic annotation is intended to allow for variation in annotation schemes while at the same time enabling comparison and evaluation, merging of different annotations, and development of common tools for creating and using annotated data. We have developed an abstract model for annotations that is capable of representing the necessary information while providing a common encoding format that can be used as a pivot for combining and comparing annotations, as well as an underlying format that can be manipulated and accessed with common tools. The details presented here provide a look "under the hood" in order to show the flexibility and representational power of the abstract scheme. However, the intention is that annotators and users of annotation schemes can continue to use their own or other formats with which they are comfortable; as long as the underlying data model is the same, translation into and out of this or any other instantiation of the abstract format will be automatic.

Our framework for linguistic annotation is built around some relatively straightforward ideas: separation of information conveyed by means of structure and information conveyed directly by specification of content categories; development of an abstract format that puts a layer of abstraction between site-specific annotation schemes and standard specifications; and creation of a Data Category Registry to provide a reference set of annotation categories. The emergence of XML and related standards, such as RDF, provides the enabling technology. We are, therefore, at a point where the creation and use of annotated data and concerns about the way it is represented can be treated separately—that is, researchers can focus on the question of *what* to encode, independent of the question of *how* to encode it. The end result should be greater coherence, consistency, and ease of use and access for linguistically annotated data.

6. Conclusion

ISO TC37/SC4 is just beginning its work, and will use the general framework discussed in the preceding sections as its starting point. However, the work of the committee will not be successful unless it is accepted by the language processing community. To ensure widespread acceptance, it is critical to involve as many representatives of the community in the development of the standards as possible, in order to ensure that all needs are addressed. This paper serves as a call for participation to the language processing community; those interested should contact the TC 37/SC 4 chairman (Laurent Romary: romary@loria.fr).

7. References

Bird, S. & Liberman, M. (2001). A formal framework for linguistic annotation. *Speech Communication,* 33:1-2, 23-60.

Bunt. H. & Romary, L. (to appear). Towards Multimodal Content Representation. *Proceedings of the Workshop on International Standards for Terminology and Language Resource Management*, Las Palmas, May 2002.

Ide, N. & Romary, L. (2001b). A Common Framework for Syntactic Annotation. *Proceedings of ACL'2001,* Toulouse, 298-305.

Ide, N., Kilgarriff, A., & Romary, L. (2000). A Formal Model of Dictionary Structure and Content. *Proceedings of Euralex 2000,* Stuttgart, 113-126.

Ide, N. & Romary, L. (2001a). Standards for Language Resources, *IRCS Workshop on Linguistic Databases*, Philadelphia, 141-49.

Ide, N. & Romary, L. (forthcoming). Encoding Syntactic Annotation. In Abeillé, A. (ed.). *Treebanks: Building and Using Syntactically Annotated Corpora.* Dordrecht: Kluwer Academic Publishers.