Department of Arts and Culture

TAMA 2003 South Africa

Terminology in Advanced Management Applications

6th INTERNATIONAL TAMA CONFERENCE: CONFERENCE PROCEEDINGS

"Multilingual Knowledge and Technology Transfer"

International Network for Terminology

TAMA 2003 South Africa

CONFERENCE PROCEEDINGS

Department of Arts and Culture

TAMA 2003 South Africa

Terminology in Advanced Management Applications

6th INTERNATIONAL TAMA CONFERENCE: CONFERENCE PROCEEDINGS

"Multilingual Knowledge and Technology Transfer"

Edited by:

Gilles-Maurice de Schryver

International Network for Terminology

Pretoria: $(SF)^2$ Press
ISBN 1-919965-01-7

Preface

This book brings together all the papers that will be presented at the *6th International TAMA Conference*, to be held in South Africa, 19-21 February 2003. All papers were edited, some even extensively. Time pressure did not allow for proofs to be sent back to the presenters. We do hope, however, that the availability of these Proceedings as *Pre*ceedings, will contribute to the success of the conference. All papers, whether meant for the pre-conference Workshop or Conference, have been placed in alphabetical order on first author, except for the keynote address which was placed first.

Financial contribution to *TAMA 2003 South Africa* by the Department of Arts and Culture, as well as by Transnet Limited, is hereby acknowledged. Thanks are also due to A. Drame and R. Koekemoer for the help in collecting the material from the presenters.

The papers were reviewed by the following Review Committee:

- Khurshid AHMAD, University of Surrey, UK
- Mariëtta ALBERTS, PanSALB, RSA
- Sonja E. BOSCH, University of South Africa, RSA
- Lynne BOWKER, Ottawa University, Canada
- Teresa CABRÉ, IULA, Spain
- Key-Sun CHOI, KORTERM, KAIST, Korea
- Georgeta CIOBANU, Universitatea Politehnica Timisoara, Romania
- Rute COSTA, Universidade Nova de Lisboa, Portugal
- Anja DRAME, TermNet, Germany
- Rika KOEKEMOER, NLS, RSA
- Christer LAURÉN, University of Vaasa, Finland
- Alan K. MELBY, Brigham Young University, USA
- Sergey PAPAEV, VNIIKI, Russia
- Laurent ROMARY, LORIA, France
- Klaus-Dirk SCHMITZ, FH Köln, Germany
- Frieda STEURS, Lessius Hogeschool, Belgium

— The Editor.

Table of Contents

Programme Pre-Conference Workshops

19 February 2003

Chairperson:	Dr. N. MGIJIMA, Chief Director: *National Language Service, SA*
09:00 – 09:15	**Welcome and Opening Address:** Mrs. B.S. MABANDLA, Deputy Minister: *Department of Science and Technology, SA*
09:15 – 10:00	• Justus ROUX • Developing Human Language Technologies in South Africa: Challenges and Proposals
10:00 – 10:30	• Anne-Marie BEUKES • Facilitating Equitable Access to Government Services through Telephone Interpreting
10:30 – 11:00	Tea break
11:00 – 11:30	• Sonja E. BOSCH, Laurette PRETORIUS & Linda VAN HUYSSTEEN • Computational Morphological Analysis as an Aid for Term Extraction
11:30 –12:00	• Jennifer DECAMP • Multilingual Technology and Technology Transfer
12:00 –12:30	• Uwe QUASTHOFF & Christian WOLFF • Web Services in Language Technology and Terminology Management
12:30 – 13:30	Lunch
Chairperson:	Dr. Gabriele SAUBERER, TermNet, Austria
13:30 – 14:00	• Gerhard B. VAN HUYSSTEEN & M.M. VAN ZAANEN • A Spellchecker for Afrikaans, Based on Morphological Analysis
14:00 – 14:30	• David JOFFE, Gilles-Maurice DE SCHRYVER & D.J. PRINSLOO • Introducing TshwaneLex – A New Computer Program for the Compilation of Dictionaries

14:30 – 15:00		• Rodmonga K. POTAPOVA & V.V. POTAPOV • Special Multiobjective and Multilingual Knowledge of Electronic Encyclopaedia
15:00 – 15:30	Tea break	
15:30 – 16:00		• Claudia BLASCHKE • Distributed Terminology Management: Modern Technologies in Client/Server Environments
16:00 – 16:30		• Ewald GEHRMANN • Case Studies on Term Entry and Glossary Distribution
	Social programme: Cocktail Party at 19:00 Dress: Smart-Casual	

Programme TAMA 2003 SA Conference

20 February 2003

Chairperson:	Prof. M. JOKWENI, University of the Western Cape, SA
08:30 – 09:00	• **Keynote Address:** Christian GALINSKI • New Horizons for Terminology Planning: e-content and m-content in the Age of the Multilingual Information Society. Standardisation of methodology concerning terminologies and language resources in support of content management
09:00 – 09:30	• Xolile T. MFAXA • Terminology Coordination and Copyright Issues
09:30 – 10:00	• Mariëtta ALBERTS • Collaboration Between PanSALB and Terminology Structures
10:00 – 10:30	• Khurshid AHMAD • Events, dear boy: Dead Cat Bounce and Falling Knife
10:30 – 11:00	Tea break
11:00 – 11:30	• Sergey PAPAEV • The Role of Terminology and Classifications for Knowledge Bases
11:30 – 12:00	• Rachélle GAUTON, Elsabé TALJARD & Gilles-Maurice DE SCHRYVER • Towards Strategies for Translating Terminology into all South African Languages: A Corpus-based Approach
12:00 – 12:30	• Nolwazi MBANANGA • Multi-cultural and Multi-lingual Society: A Challenge for e-Health in South Africa
12:30 – 13:00	• Xinli YU & Min SONG • The Development of Computer-aided Term Extraction Software

13:00 – 14:00	Lunch

Chairperson:	Prof. M.J. MOJALEFA, University of Pretoria, SA
14:00 – 14:30	• Maria SMIT • E-Learning in Music: Insights Gained from the Compilation of an Electronic Database on African Music Terms
14:30 – 15:00	• Nonkosi TYOLWANA • The Development of Terminologies in African Languages as a Key to Sustainable Human Development and Empowerment
15:00 – 15:30	• Michele VAN DER MERWE • Towards the Creation of a Dictionary Culture in South Africa

15:30 – 16:00	Tea break

16:00 – 16:30	• D.J. PRINSLOO & Gilles-Maurice DE SCHRYVER • Towards Second-Generation Spellcheckers for the South African Languages

	Social programme: Banquet at the Indaba Hotel at 19:00 Address by the Minister of the *Department of Arts and Culture, SA* Dress: Dark Suit

Programme TAMA 2003 SA Conference

21 February 2003

Chairperson:	Prof. N.C.P. GOLELE, PanSALB, SA
07:30 – 08:00	• Zola WABABA, T. MBATHA & B. MAHLALELA • A Terminology Development Initiative (Corpus Planning from Below) in a Dual-Medium Science Project of PRAESA
08:00 – 08:30	• Mark D. CHILDRESS • The Practical Use of Knowledge Management Theory in Terminology Management
08:30 – 09:00	• W. VAN ZYL DE VILLIERS • The Compilation of a Quadrilingual Explanatory Dictionary of Chemistry
09:00 – 09:30	• A. AKINYEMI • Language Factor in e-Learning for Technology Transfer
09:30 – 10:00	Tea break
10:00 – 10:30	• Johan D.U. GELDENHUYS • Terminological Shift in a Slippery Economy
10:30 – 11:00	• Bassey E. ANTIA & André CLAS • Terminology Evaluation
11:00 – 11:30	• Barbara I. KARSCH • The Evolution of Version 2 – A Multilingual Database for A Multitude of Users
11:30 – 12:00	• Klaus-Dirk SCHMITZ • DTP – the German Terminology Portal
12:00 – 12:15	Break
Chairperson:	Prof. R. FINLAYSON, University of South Africa, SA
12:15 – 12:45	• Frieda STEURS • Translation Technology and Workflow Procedures in Technical Documentation Management

12:45 – 13:15	• Irina N. ROZINA, Ronald D. ECKARD & Joe DOWNING • Asynchronous Learning Environments for Language, Communication and Culture Study
13:15 – 13:45	• Christian GALINSKI • The Way Forward
13:45 – 14:00	**Closure:** Representative of the *Department of Arts and Culture, SA*

14:00 –	Lunch

New Horizons for Terminology Planning: e-content and m-content in the Age of the Multilingual Information Society

Standardisation of methodology concerning terminologies and language resources in support of content management

Christian GALINSKI
TermNet, Austria

Abstract: Recently more and more aspects of the 'economics of language' (viz. primarily the costs of the use of language in specialised / professional communication) in general and of terminology in particular are identified. As communication consumes time or causes transaction 'expenses' in some way or other, costs are incurred every time we are communicating. Some of these financial or non-financial 'expenses' are not yet measurable, other have become measurable. This applies to inter-personal communication by 'natural language' – whether in oral form or in written form, whether in general purpose language (GPL) or in special purpose language (SPL) –, and to man-machine communication, as well as to communication by language between computers. *Of course the objective is not to avoid communication, but to render communication more efficient and effective at places, in environments, at times, where and when it is necessary or useful*.

Here methodology unification / standardisation / harmonisation provides important clues for cost reduction, and at the same time for the improved quality of communication. This refers in particular to the unification / standardisation / harmonisation of methodologies concerning the preparation, processing and use as well as re-use of terminologies and language resources (TLRs) in support of content management, and also refers to the respective metadata for the sake of re-usability of data as well as to data structures for the sake of interoperability between different data models. Therefore, the Technical Committee ISO / TC 37 "Terminology and other language resources" of the International Organisation for Standardisation (ISO) has opened its scope towards language resources in general during the last couple of years. This was due, among others, to the following considerations:

- terminology is in most cases embedded in or combined with LRs;
- new developments in the information and communication technologies (ICTs) – especially in mobile computing and mobile communication (MCC) applied to mobile content, mobile commerce, etc. – increasingly require the integration or combination of all kinds of content (incl. TLRs);
- TLRs increasingly have to be treated as multilingual, multimedia and multimodal from the outset.

It is acknowledged by now in industry that "products have to be marketed in the language of the target markets". Product here also comprises services, and the accompanying documentation (in the meaning of product description, manuals, technical handbooks, etc.) is an integral part of the product (in line also with the quality management approach). This also applies to 'intangible products' – like information products and services – which can be regarded as a 'commodity' in professional and social life in the global information society.

Major mobile telephone companies (telcos) and MT (mobile telephony) service providers have recognised that the further development of business via MCC (mobile computing and mobile communication extending towards e-business, m-commerce, etc.) is based on three pillars:

- content;
- technology;
- business models.

For businesses based on content, there are three key success factors, namely appropriate solutions for:

- the efficient use of language (incl. human language technologies (HLTs) and, of course, multilingual terminologies and LRs);
- existence of standards (especially methodology standards referring to multilinguality, metadata, data modelling and XML applications);
- transfers (of content first of all, but also concerning broadband access, micro-payment systems, etc.).

This was clearly voiced by MT company CEOs at the last MOST Conference (organized by the Think Tank of the Initiative "Mobile Open Society through Telecommunication") in Warsaw, October 2002.

1. Multilingual aspects of the global information society (GIS)

As soon as a given language has passed the basic stages of language planning (i.e. when the 'linguistic norm' is fixed by establishing a standard orthography and grammar), terminology planning should set in. Without terminology planning the language cannot develop into a tool for professional (or specialised) communication or might be reduced to its GPL (general purpose language) role in folklore and local / regional culture. Terminology planning differs from language planning with respect to other conventions of SPL (special purpose language) use and the 'creation' or adoption of terminologies, but should follow as much as possible the basic rules of the GPL in question. In any case a systematic approach with certain normative guidelines should be conceived, also including rules for term formation or borrowing.

Humans communicate in order to exchange ideas, transfer knowledge, hand down culture, express feelings, etc. which to a large extent occurs via content. Content here means any semiotic representation of information and knowledge. It can take the form of TLRs or non-linguistic representations (such as graphical information, etc.), which can increasingly be processed by the computer. Information and knowledge management cannot work effectively without proper content in the form of linguistic and non-linguistic knowledge representations. In future cyberspace the availability of re-usable and interoperable language resources is of utmost importance not only for the dialogue between all kinds of communities but more and more also for industrial and commercial activities.

ICTs provide the technical infrastructure and tools to support inter-human communication as well as the processes to create, process, disseminate and re-use 'content' (which is primarily representing knowledge) for multiple purposes and

applications. Increasingly all processes of the creation, processing, dissemination and re-use of content are influenced by 'wireless' (mobile) applications of the ICTs. Contrary to some people's impression the globalising forces of the ICTs do not necessarily curb cultural diversity. New cultural forms emerge, however; cultural change is accelerated. In this process languages and cultures are competing at global level, with fair chances for small language communities not only to survive, but also to develop – *if they make the necessary efforts*!

The ICT infrastructures are global, their use, however, is local. Therefore, ICTs – also being 'products' and the respective services as well as content – increasingly need to be 'localised'. Localisation is the process of adapting products and services to a specific local environment, involving the use of appropriate character sets, translations and other aspects that make the products and services usable for users in that specific culture. So multilinguality and cultural diversity (MCD) have to be taken into account. Localisation is most efficient, if it can build on internationalisation. Internationalisation is the process whereby products and services are implemented in a way that allows for and facilitates the adaptation to local languages and cultural conventions (i.e. MCD). Internationalisation is a prerequisite for a systematic and thus efficient approach to localisation.

As stated above *multilinguality* has to been seen in the wider perspective of *multilinguality and cultural diversity* (MCD), which has an increasing impact on *cultural adaptability*. The discussion has started at international level in the UN framework – especially at the UNESCO – and has extended to the European level (see Matteini 2001). The ground was prepared by the European Programmes MLAP (Multilingual Action Plan), EAGLES (Expert Advisory Group on Language Engineering Standards – followed by the Programme ISLE, International Standards for Language Engineering) and MLIS (Multilingual Information Society) in the past, and is continued by today's e-Content Programme. Since a couple of years the discussion has reached management level in big industry.

Multilinguality is quickly becoming a major issue for the European telecos, which are developing into full-service companies, offering a wide range of services (e.g. in the form of e-commerce) via the Internet. Many Web services must be offered in several languages. The design and maintenance of multilingual websites require tools and procedures well beyond what is needed for monolingual websites. Without suitable tools – based on standardised architectures for multilingual websites – these sites and the attendant services are very expensive to create and manage. ICTs are getting cheaper, content more accessible every day. Knowledge transfer, therefore, could, if properly supported, be largely facilitated.

2. Standardisation

From the above it becomes clear that MCD has an impact on ICTs (in terms of both hardware and software), content and the methodologies to create, process and maintain

content, as well as on human behaviour. This has been acknowledged in standardisation in the form of cultural adaptability, which is defined by ISO / IEC JTC 1 as *'the special characteristics of natural languages and the commonly accepted rules for their use (especially in written form) which are particular to a society or geographic area. Examples are: national characters and associated elements (such as hyphens, dashes, and punctuation marks), correct transformation of characters, dates and measures, sorting and searching rules, coding of national entities (such as country and currency codes), presentation of telephone numbers, and keyboard layouts'*. Cultural adaptability is closely related to processes such as globalisation, internationalisation, localisation, and to some degree personalisation too.

In this connection language-independent approaches applied to content management methods have proven to be most effective. They avoid language-pair comparison / translation / conversion as much as possible for the sake of highest efficiency and effectiveness as well as cost saving in various applications, such as data modelling, localisation, etc. based on multilingual content management in combination with the appropriate HLT (human language technology) tools. In fact this means methodology standardisation in contrast to technical standards focused on ICTs. Methodology unification / standardisation / harmonisation provides important clues for cost reduction, and at the same time for the improved quality of communication. This refers in particular to the methods concerning language resources (LRs) for the sake of content management, and may refer to the data themselves as well as to data modelling.

Today the metadata approach is state-of-the-art for linking and evaluating information on the Web by making it interoperable. By means of metadata – in the meaning of identified, formally described data elements – the problems of multilinguality of TLRs can be solved comparatively easily. Therefore, ISO / TC 37 has adopted the metadata approach in its HLT related activities. Language resources, such as written and spoken corpora, computational lexicons, terminology databases, speech collection and processing, etc. can be defined as a set of speech or language data and descriptions in machine readable form, used e.g. for building, improving or evaluating natural language and speech algorithms or systems, or as core resources for the software localisation and language services industries, language studies, electronic publishing, international transactions, subject-area specialists and end users. The metadata approach also is a prerequisite for interoperability, i.e. the achievement of partial or total compatibility between heterogeneous data models by mapping of metadata.

At present the creation of those kinds of content, which are based on LRs, is still too slow, too expensive, mostly not good enough and rarely with a guarantee for correctness. ISO / TC 37 is trying to improve this development by preparing standards and other documents with rules as well as guidelines for:

- harmonised metadata;

- unified principles and methods for data modelling;
- standardised meta-models;
- workflow management methods for net-based distributed cooperative creation of terminology and other language resources.

This kind of methodology standardisation not only enhances the performance of content creation, but also ensures the re-usability of data (for other environments, other purposes, different uses and over time) as well as interoperability of data structures. This in fact decreases costs dramatically.

3. Content management

Increasingly, system designers and developers recognise that more refined data models (in terms of a higher granularity and a higher level of international unification and harmonisation) can enable information and knowledge management in the organisation to cope with the above-mentioned cost situation. A higher degree of standardisation of methodology with respect to TLRs, is a prerequisite for achieving satisfactory solutions for information and knowledge management based on multilingual content management in the enterprise. This was thoroughly investigated by Martin (2001).

E-business – especially in combination with mobile computing resulting in m-commerce – is probably going to change the organisation and operation of enterprises and their business quite radically in the near future. Enterprises and other organisations / institutions will be forced not only to link hitherto separated systems to each other, but to really 'integrate' all data processing systems of the organisation – of course including their content. Latest at this point, the whole degree of variation in language usage within the organisation will become apparent.

Industry and trade are already preparing for applying language engineering methods and tools to lesser-used languages down to the dialect level. Therefore, lesser-used languages (often also: minority languages) can benefit from new chances for development in the emerging multilingual information society – if, again, the respective language communities are prepared to do the necessary efforts.

4. e-Content

A recent study for the European Commission (Andersen 2002) identifies, among others, the following transaction-centric and content-centric kinds of m-content or m-content services which are already emerging (left hand side), to which enhanced future kinds of m-content or m-content services could be added (right hand side):

Content based on language resources →	***multilingual, multimedia, multimodal* m-content**
• mobile general news • mobile transport information	• mobile public information • mobile transport and delivery information

• mobile financial data	• mobile financial services
• mobile games	• MT services for professionals
• mobile edutainment	• mobile learning and training
• mobile music	• mobile composing
• mobile transaction services	• mobile B2B services
• mobile directories	• mobile directories for professionals
• mobile adult information	• mobile information for professionals

It seems that MCC will further drive the need for multilingual TLRs and the respective methods as well as for methodology standardisation.

For enterprises using the Internet for e-commerce the general principle "sell products / services globally, but market them locally" applies – even, if they may not recognise it at the beginning. ICTs are changing nearly everything in society – even language as such and the application of languages as well as the cooperation of people using language. Multilinguality applies to nearly every aspect of an information system:

- language use in general;
- content in terms of:
  - terminologies,
  - language resources;
- access to information;
- special adaptations for disabled persons;
- cyberspace: complex network of networks developing out of the Internet and other information networks;
- networking;
- communication;
- knowledge and knowledge databases;
- understanding and intercultural communication.

In this connection intercultural aspects have more influence on data modelling and programming than one might expect. This and the respective needs for standards as well as future requirements have been extensively investigated in a number of reports.

In a letter to *Business Week* (April 8, 2002) Berners-Lee (MIT, the 'father' of the "Semantic Web" conception) denies that the WWW will be replaced by the Semantic Web, with the following arguments:

> The WWW contains documents intended ***for human consumption***, and those intended ***for machine processing***. The Semantic Web will enhance the latter. The Semantic Web will not understand human language ... The Semantic Web is about machine languages*: well-defined, mathematical, boring, but processable. ***Data, not poetry***.
> *[* in the meaning of highly standardised natural language or highly controlled language]*

Berners-Lee thus indicates that he is widely misunderstood or misinterpreted.

These remarks also point in the direction of how language use in the information and knowledge society in general and in future e-business (comprising the whole range of e-commerce, e-procurement, e-content, etc. to m-commerce) will develop: highly harmonised terminology combined with factual data and common language elements need to be provided in a form:

- presumably nearer to human natural language usage in B2C;
- presumably nearer to highly 'controlled languages' in B2B.

What is new in this connection is that these machine languages will also be multilingual in terms of human language use. Beside, they will be multilingual, multimodal and multimedia from the outset.

5. Conclusion

Hardware costs are not only decreasing year by year, hardware components are also nearing the time-honoured ideal of 'plug-and-play' according to the OSI standard (open system interconnection). Software still is far too expensive – not in terms of purchase, but in terms of the necessary adaptation and continuous upgrading. In addition software today is still far away from 'plug-and-play'. But the increased emergence of open-source software will reduce costs in the long run. High-quality content creation and maintenance, however, still is and will be the biggest cost factor. In analogy to ISO's OSI model we need something like an OCI (Open Content Interoperability) model – which in fact is the vision of ISO / TC 37.

A higher degree of standardisation of methodology with respect to TLRs, is a prerequisite for achieving satisfactory solutions for information and knowledge management based on content management in the enterprise. Increasingly system designers and developers recognise that only:

- more refined data models (in terms of a higher granularity and a higher degree of international unification and harmonisation),
- the application of standards, and
- the application of the appropriate methodologies,

can enable content management in the organisation to cope with the array of problems posed by accelerated globalisation – and the need for more localisation in its wake.

References & Bibliography

Andersen (ed.). 2002. *Digital content for global mobile services. Final report*. Luxembourg: CEC.

Berners-Lee, T. 2002. Reader's letter. In *Business Week*, 8 April 2002.

CEN (ed.). 1999. *Model for metadata for multimedia information*. Brussels: CEN. (CEN / ISSS / CWA 13699:1999)

CEN (ed.). 2000. *Guidance information for the use of the Dublin Core in Europe*. Brussels: CEN. (CEN / ISSS / CWA 13988:2000E)

CEN (ed.). 2000. *Dublin Core metadata element set*. Brussels: CEN. (CEN / ISSS / CWA 13874:2000)

CEN (ed.). 2000. *Description of structure and maintenance of the Web based Observatory of European work on metadata*. Brussels: CEN. (CEN / ISSS / CWA 13989:2000)

CEN (ed.). 2000. *Information Technology – Multilingual European subsets in ISO / IEC 10646-1*. Brussels: CEN. (CEN / ISSS / CWA 13873:2000)

CEN (ed.). 2001. *European culturally specific ICT requirements*. Brussels: CEN. (CEN / ISSS / CWA 14094:2001)

Clews, J. and H. Hjulstad (project team). 2002. *Programming for cultural diversity in ICT systems. A Project Team report to CEN / ISSS on consensus-building in European standardisation. Final version 2002-09-23*. Brussels: CEN.

EURESCOM Report on P923 "Multilingual web sites: Best practice, guidelines and architectures. D1 Guidelines for building multilingual web sites – Sept. 2000"

GAO (ed.). 2002. *Electronic government. Challenges to effective adoption of the Extensible Markup Language. Report to the Chairman, Committee on Government Affairs, U.S. Senate*. Washington: United States General Accounting Office. (GAO-02-327)

Hawkins, R. (ed.). 2000. *Study of the standards-related information requirements of users in the information society*. Brussels: SPRU. (Final Report to CEN / ISSS 14 February 2000)

Hovy, E. et al. (eds.). *Multilingual information management: Current levels and future abilities. A report commissioned by the US National Science Foundation and also delivered to the European Commission's Language Engineering Office and the US Defence Advanced Research Projects Agency. April 1999*. (Web version 23/10/2000: <http://www2.hltcentral.org/hlt/download/MLIM.html>)

Martin, B. 2001. Terminology management driving content management. In F. Steurs (ed.). *TAMA 2001. Terminology in Advanced Microcomputer Applications. Sharing terminological knowledge. Terminology for multilingual content*: 26-39. Vienna: TermNet Publisher.

Matteini, S. 2001. *Multilinguality and the Internet*. European Parliament. (Research Directorate A. STOA – Scientifical and Technological Options Assessment. Briefing Note No. 2 / 2001). <http://www.europarl.eu.int/stoa/publi/pdf/briefings/02_en.pdf>

Pricewaterhouse, C. (ed.). 2001. *Cultural diversity market study. Final Report*. Luxembourg.

UNESCO (ed.). 2001. *Report of the General Conference 31st Session*. Paris, 15 October to 3 November 2001. Document 31 / C5. Volume 1 Resolutions: 68-69.

UNESCO (ed.). 2001. *Draft recommendation concerning the promotion and use of multilingualism and universal access to cyberspace*. Paris: UNESCO. (Document C31 / 25 Corr.)

Appendix 1
Traditional and new content creation and data modelling

Traditional Data Modelling	**Enhanced Data Modelling**
• mono-purpose	• multi-purpose and multi-functional
• textual data	• graphical symbols, formula, etc. • images and other visual representations • multimedia, multimodal
• TLRs data categories	• additional ontology data categories • higher degree of granularity
• data elements repeatable by language and within language	• other kinds of repeatability (e.g. register) • qualifiers, attributes, properties, etc. • statistics, validation, copyright management, etc.
• language independent approach	• multimedia and multimodality (incl. non-linguistic representations)
• by subject-field experts, LR experts with subject-field expertise	• by anybody according to level of expertise... (→ sophisticated access right management)
• traditional systematic / semi-systematic approach	• using also other kinds of systematic approaches
• conventional DB management	• sophisticated database management methodology (based on metadata approaches for distributed DBs)
• conventional quality control	• automatic validation, maintenance, copyright management

Traditional Content Creation	**New Methods of Content Creation**
• by one subject-field expert • by one LR expert	LR expert can serve as: • consultant • project manager
• by a group of experts (subject-field experts or specialised LR expert with subject-field expertise or mixed	• net-based distributed co-operative work to establish content databases • including terminological and other

composition of expert group: ▪ majority experts with the assistance of one or a few terminologists; ▪ majority terminologists with the assistance of one or a few experts).	language / knowledge resources • additional features: ▪ (semi-)automatic validation ▪ copyright management → all users are potential creators of data → economies of scale in contents creation

Appendix 2

ISO / TC 37 "Terminology and other language resources"

(PWI "Basic principles of multilingual product classification for e-commerce")

ISO / TC 37 / SC 1 "Principles and methods"

- **WG 2** "Vocabulary of terminology"
- **WG 3** "Principles, methods and concept systems"
- **WG 4 "Terminology of socio-linguistic applications"*

ISO / TC 37 / SC 2 "Terminography and lexicography"

- **WG 1** "Language coding"
- **WG 2** "Terminography"
- **WG 3** "Lexicography"
- **WG 4** "Source identification for language resources"

ISO / TC 37 / SC 3 "Computer applications in terminology"

- **WG 1** "Data elements"
- **WG 2** "Vocabulary"
- **WG 3** "Data interchange"
- **WG 4** "Database management"

ISO / TC 37 / SC 4 "Language resource management"

- **WG 1** "Basic descriptors and mechanisms for language resources"
- **WG 2 "Representation schemes"*
- **WG 3 "Multilingual text representation"*
- **WG 4 "Lexical database"*
- **WG 5 "Workflow of language resource management"*

**planned*

Events, dear boy[*1]: Dead Cat Bounce and Falling Knife

Khurshid AHMAD
Department of Computing, University of Surrey, UK

1. Preamble

The two terms in the title of this paper, *Dead Cat Bounce* and *Falling Knife* are popularly used in financial reports, especially those concerning the trading of specific financial instruments (an instrument is a super-ordinate term and its instances include currencies (e.g. $, £, ¥), shares, government and private bonds). The value of these instruments is established by the market conditions in which they are traded and are therefore determined, in significant measure, by *market sentiment*. *Dead Cat Bounce*, a metaphorical term, refers to a temporary recovery by a market after a prolonged decline or bear market. In most cases the recovery is temporary and the market will continue to fall. The term *Falling Knife* refers to a stock whose price is in the middle of a big fall from a previous value.

A whole range of metaphors is utilised in writing about or discussing financial matters, for example, the animal metaphors *bear* and *bull* markets; or health metaphors such as *anaemic* currencies / economies; or spatial metaphors – instruments go *up, down*, instruments *crash*, find their own *level*. One can argue that market sentiments are determined, in some measure, by *events*: a set of happenings that occur within a well-defined spatial confine – a small earthquake in Chile to World War II may be two examples of how inclusive the space can be – and that these happenings cover a certain extent in time – again events may last a few seconds or many years or even millennia.

When we describe an event, we describe the abstract, including ideas, simplifications, aspirations and beliefs, and we describe the concrete, especially objects, people, places. The abstract and the concrete are described in the context of a significant occurrence, happening or phenomenon.

This paper deals with some keywords related to the widely used word *event* and how one can devise a method that will help in understanding these occurrences when analysing them via news reports and other documents. Because computers are fairly unintelligent devices and are used to analyse the texts, my research method has to be simple. The news reports and documents that interest me most are written in specialist language, the constrained nature of which benefits the method of analysis.

---

[*] This paper is based on presentations the author has made at two recent workshops. The first was the 'Event Modelling for Multilingual Document Linking', LREC 2002 Workshop, Las Palmas, Canary Islands (June 2002). The second was the workshop on 'Financial News Analysis', TKE 2002, Nancy, France (August 2002).

[1] Attributed to a former British Prime Minister, the late Harold Macmillan, who, when asked what can scupper the best-laid plans of a politician, was reported to have said 'Events, dear boy'.

The constraints of special language, including the repeated use of certain preferred terms and the use of fewer syntactic structures, and the relatively simple assumptions about the organisation of knowledge within a specialist domain, are exploitable in analysing events and their causes. The assertion that special language relies on 'simple assumptions' may appear provocative and/or naïve, but when one looks closely at any 20th century description of most specialist enterprises, ranging from relativity theory through to media studies, and from engineering sciences to sociology, politics, and even philosophy, one sees intelligent men and women drawing taxonomies and hierarchies of one type or the other to describe the microscopic world, the entire Universe, kinship and exchange relationships, world trade or family disputes.

The ontological commitments of any specialist community are there for all to see: the specialist endeavour to wrap the world at large or beings in the real world or virtual worlds in taxonomies, equations, diagrams, constants and principles (including the rather pompous *universal constants and principles*). Another manifestation of ontological commitments within a specialism is its terms. The concrete and abstract that demarcate and distinguish one event from another, or indeed connect one event to another, are articulated in the description of the event through the use of certain terms and through the use of key verbs specific to that specialism.

The other key attribute of the abstract and the concrete is that their referent may be a unique idea, person, place or thing. This reference is articulated as a *proper noun*. The arbitrary nature of proper nouns – for example, many of them are *given* names – makes it difficult to understand their contribution to the description of events.

2. Describing how events are described

Current work in artificial intelligence, a branch of computing, deals with the representation of (specialist) knowledge: representation that requires a set of conventions about how to describe a class of the abstract or concrete. The notion of representation is closely intertwined, with open questions in philosophy and cognate subjects, to the notions of *meaning*, *intention* and other open-ended conundra. Schank and colleagues have been at the forefront of this ambitious enterprise and have attempted to present methods for building computer programs that can summarize a collection of sentences, programs to answer questions about the content of the sentences, and perhaps eventually to translate the collection from one language to another (see Hardt (1992) for a review of conceptual dependency).

Schank & Abelson's (1977) *Conceptual Dependency Theory* (CDT) was developed as part of a natural language comprehension project and can perhaps be regarded as one of the precursors to the debate on whether event-structure formation contains different structure information or whether this information is part of a more general conceptual or logical semantic representation. CDT has succeeded where many other theories have not quite and has been applied to early virtual reality systems. CDT

can represent action: the staple of virtual reality systems are things moving, objects colliding, people communicating, and objects and people in various states of being.

Schank's claim was that sentences could be translated into basic concepts expressed as a small set of semantic primitives. Conceptual dependency allows these primitives, which signify meanings, to be combined to represent more complex meanings. Schank calls the meaning propositions underlying language "conceptualisations". The conceptualisations can be either active or passive; the former comprise *actors, actions, objects, source, destination.* The stative conceptualisations, through an arbitrary scale ranging from –10 to +10 can indicate state changes. The stative conceptualisations of *health*, *anticipation*, *mental* and *physical states*, and *awareness* have been 'computed', that is, a computer program has attempted to infer the 'meaning' of an underlying proposition, by interpreting the scales. The statement *Bill shot Bob in the heart repeatedly until Bob was no more* will be interpreted by CDT as 'Bob: State → Health ≡ -10'; or *John thought Mary found discussions about meaning make her unhappy* which the CDT will compute as 'Mary: State → Mental State ≡ -5'. The world of action-oriented abstracts and concretes is a complex one and CDT approached it bravely. We intend to follow this approach and will attempt to focus on how to infer meaning or intent from examining the lexical content of a sentence or a collection of sentences.

In order to learn about the state of the contents, we have adopted a corpus-based approach: instead of relying on postulates about meaning, encoded as rules of syntax and semantics, we rely on the evidence based almost entirely on the frequency of lexical items. For us, frequency correlates with acceptability. For instance if there is only one instance of *John thought Mary found…* in a corpus of 100,000 sentences then, statistically, whatever *John thought* about *Mary* is in the realm of statistical outliers. Any inference drawn from outliers has to be heavily qualified. However, if the frequency of the construct *John thought Mary…* is, say, 1 in 1000 sentences, then it would be safer to infer on the basis of this sample than the one previously discussed.

3. Semantics of Events?

Some authors postulate a distinct and separate level of representation for event structure (Pustejovsky 1991) adopting the view that event structure information concerning time, space and causation has a different status from other kinds of thematic, conceptual or lexical information. Other authors assume that event structure information is part of, or is implicit in, a more general conceptual or logical semantic representation (Jackendoff 1990).

Pustejovsky (2000) has noted that *'there has been a renewed interest in the explicit modelling of events in the semantics of natural language'*. Events in this kind of work *'are associated with the tensed matrix verb of a sentence and sometimes with event-denoting nominal expressions, such as* war *and* arrival' (Pustejovsky 2000: 445). Here the claim is that if we had a well-developed system through which we can process

lexical semantic relations and a good grammatical description of how nouns behave, then we can describe "events as grammatical objects" (Tenny & Pustejovsky 2000).

Lakoff & Johnson (1999) deal at length with causation. For them, states are locations, whereby one can say things like *the Japanese economy is out of depression and the US economy is in deep depression*. Then there is a discussion of 'changes': *'a change of state [is a] movement from one bounded region in space to another bounded region in space'* (2000: 183). The description of such bounded movement involves the verbs and prepositions of motion like *go*, *come*, *enter*, *from*, *to*, *into* and *between*. The changes could be continuous or graded. So the financial metaphor would be *the revitalisation of a company or an economy*, i.e., from a state of poor health to a state of relative well-being. This so-called location event-structure metaphor involves forces that are responsible for causes and force movement which affects causation. Note that the notion of force that in our times is related almost exclusively to concrete objects is now being applied to fairly abstract concepts like an economy, or the state of an organisation.

Knowles (1996) has noted that financial journalists are keen on health metaphors – *anaemic*, *ailing*, *debilitating*, *fatal*, *feverish*, *haemorrhaging* – to describe a failing economy, or a falling currency, or a poorly performing bond or a crashing share. And, when a financial instrument is buoyant the journalists appear to celebrate the well-being of the market by using metaphorical terms like *immunity*, *revitalisation*, *appetite*, *strength* and so on.

The works of Lakoff & Johnson, of Pustejovsky, and the empirical observations of Knowles allow us to build a framework for analysing financial news stories, reports and learned articles. Metaphors will pose considerable challenges to the current systems for information extraction that deal with news stories, reports and learned articles in finance.

4. Events and sentiments: A case study

Our work focuses on movements of financial instruments including weighted indices of national stock exchanges like the UK Financial Times (London) Stock Exchange 100 (top companies) Index, better known as the FTSE and pronounced *footsie*. We report on some initial work that attempts to compare changes in the FTSE100 with changes in 'market sentiment' as expressed in news reports about the UK economy specifically, and reports about Wall Street indices. The latter has a substantial influence on the UK economy. In addition to sophisticated metaphors there are a number of verbs and adjectives that describe the sentiment of traders with respect to the market they trade in. There are fairly literal words that express sentiment, as reported in the news wires, about the markets: financial instruments *rise, fall,* markets *boom, go bust,* and there are *gains, losses* within the markets, economies *slow down*, suffer *downturns*, whole industry sectors may be *hard pressed*. Below are some examples of news that may express good (or positive) sentiment and bad (or negative) sentiment:

Date	Time	Left context	Positive sentiment	Right context
17 Sep 2002	09:54	insurance premiums, no passenger	**growth**	in uk trains and fare
17 Sep 2002	11:58	tesco's	**growth**	has slowed from seven percent
18 Sep 2002	09:22	said, with the greatest	**growth**	in widebody freighters such as
25 Sep 2002	10:59	but smiths said it expected	**growth**	in military aerospace, medical
27 Sep 2002	10:56	see some acceleration in output	**growth,**	particularly in services

Date	Time	Left context	Negative sentiment	Right context
02 Sep 2002	15:54	consumer spending intentions	**fell**	five points to - 8
03 Sep 2002	16:52	source of future growth,	**fell**	for the first time since
05 Sep 2002	13:12	percent, while prudential	**fell**	five percent after it said
09 Sep 2002	12:46	in troubled music firm emi	**fell**	five percent to 159 -
13 Sep 2002	17:28	consumer sentiment	**fell**	for a fourth straight month

These news reports are written in free natural language and we expected, and found, some 'misleading' sentences like *New Zealand captain Stephen Fleming (79 not out) and debutant opener Lou Vincent (86 not out) reached their half-centuries in an unbeaten 171-run partnership, adding 103 runs without* ***loss*** *for the middle session after resuming at 87 for two.* Here the register has changed from finance to sports and this we endeavour to guard against.

Here are the results of an analysis carried out automatically on 980 financial news items published by Reuters daily (except Saturday and Sunday) in November 2002 comprising over 400,000 words. Our system, based on System Quirk, analysed 70 sentiment words, divided equally between positive and negative sentiment expressing words, in this news corpus. The frequency of positive sentiment words (normalised values starting at c. 0.2) are plotted with the closing value of the FTSE 100 for the month of November 2002: The correlation is encouraging:

5. Afterword and work in progress

The above analysis and the concomitant results are of a tentative nature in that work is progressing in four major directions. First, we are looking at how accurately our chosen single sentiment word can convey market sentiment; initial results are encouraging in that few of the 70 sentiment words are used most frequently and that these frequently used sentiment words occur in a restricted set of phrase structures. Second, we are exploring the use of financial terminology, based on a commercial Web-based financial terminology system (http://www.investorwords.com), for categorising the financial news into various different categories – stocks, currencies, investment banking – with a view of determining sentiments relating to a given instrument. Third, it has been estimated that over 30% of a financial news item comprises proper nouns – names of organisations and people – and an identification of the proper nouns may lead to the attribution of sentiment to (a group of) people or organisations. Fourth, perhaps most importantly, we have described a *time series* relating to sentiments: usually, a time series is about cardinal numbers related to a concrete entity – temperature, pressure, money supply, units of goods produced – and are measured using instruments of one kind or another or merely counted by human beings. It is true that opinion polls about politicians or political parties, impart sentiments expressed by people, and to quantify the expression of such opinions over time is a novel time-series construct. We are in the process of investigating the extent to which we can assemble a time series on the basis of indirect observations and what it means to correlate such a series with other more quantitative series like the FTSE100.

Ours is a deliberately lexical approach to making inferences from texts. We are guided by major work in knowledge representation, especially the representation of events, and work in semantics and lexical semantics. Computers currently can process lexical information with some success. It can be argued that the more recent developments in the study of language, both special and general language, and statistically-oriented corpus linguistics are entirely dependent on the abilities of computers to store and (repeatedly) search for lexical patterns in large volumes of texts. The study of how events can be described can benefit automatic extraction of information, both theoretically and practically, from a lexical, corpus-based approach.

Acknowledgements

The work was carried out under the partial sponsorship of the EU's 5th Framework Programme on Information Societies project GIDA (Project No. IST 2000-31123). The research team at the University of Surrey is led by Khurshid Ahmad and includes Saif Ahmed, David Cheng, Tony Chiu, Pensiri Manomaisupat, Paulo C.F. de Oliveira, and Tugba Taskaya. Lee Gillam is the Technical Manager of the GIDA project at Surrey and Matthew Casey is the Research Supervisor.

References

Hardt, S.L. 1992. Conceptual Dependency. In S.S. Shapiro (ed.). *Encyclopaedia of Artificial Intelligence*: 259-262.

Jackendoff, R. 1990. *Semantic Structures*. Cambridge: MIT Press.

Knowles, F. 1996. Lexicographical Aspects of Health Metaphors in Financial Texts. In M. Gellerstam et al. (eds.). *Euralex'96 Proceedings (Part II)*: 789-796. Gothenburg: Göteborg University.

Lakoff, G. and M. Johnson. 1999. *Philosophy in the Flesh*. New York: Basic Books.

Pustejovsky, J. 1991. The Syntax of Event Structure. *Cognition* 41: 47-81.

Pustejovsky, J. 2000. Events and the Semantics of Opposition. In C. Tenny and J. Pustejovsky: 445-482.

Schank, R.C. and R.P. Abelson. 1977. *Scripts, Plans, Goals and Understanding*. Hillsdale: Erlbaum.

Tenny, C. and J. Pustejovsky (eds.). 2000. *Events as Grammatical Objects*. Stanford: CSLI Publications.

Language Factor in e-Learning for Technology Transfer

A. AKINYEMI
Center for Educational Technology, Sultan Qaboos University, Sultanate of Oman

Abstract: Information and communication technologies have permeated the way people of all nations live. Technology transfer had been the popular terminology used for the developing countries for a long time. Today computer technologies have webbed the world together as one global village thus making technology transfer unavoidable and inevitable. E-learning has become the popular mode of instruction in business and in higher education and most of all a veritable instrument for change. Language is an important aspect of communication, education and learning which cannot be separated from forms of interactions in the learning process. Electronic-learning, which may be viewed as a virtual teacher, pseudo-teacher or complementary teacher, is characterised by forms of learner interactions. The study of language reveals its importance, whether as first, second or tenth and these are crucial in the conduct of e-learning. In as much as interaction is being orchestrated by an invisible tutor, language becomes the hallmark of this new approach. Language must have some significant roles in the way in which learning occurs and consequently, in the success or failure of e-learning for an effective technology transfer from one setting to another in the global village.

1. Communication technologies

Information and Communication Technologies (ICT) have taken over the educational scene like a wild fire. The introduction and rapid spread of ICT have revolutionised the way societies interact, conduct business and the contemporary practice of teaching and learning. The World-Wide Web (Internet), e-mail, telephones (cellular technology), digital technologies (CD-ROM), fibre optics, satellites and many technologies have changed the way we live in the new Millennium. ICT has been further realised to be important in the effort to eradicate poverty in poor nations. No one is in doubt that technology is important in every aspect of our lives, however, it cannot be regarded as a panacea in teaching and learning. In the words of Goldberg:

> Technology cannot change who we are or the way we treat our students, but it helps connect us with those at a distance and puts us more in touch with the activities and thoughts of all our students. (Murray Goldberg, Founder of WebCT)

ICT is an important aspect of growth and development of a poor nation in more ways than in the conduct of education. It can contribute towards strengthening democracy, increasing social participation, competing in the global market place and removing barriers to modernisation as it makes poor populations fuller agents in the sustainable developmental process. Before the full benefits of ICT can be realised, poor nations may need to pay more attention to the convergence of the new and traditional communication media in view of the dearth of requisite infrastructures to sustain the new technologies. ICT is not just about technologies, its importance is further realised

in information transfer and communication in many domains. All nations need to be vast in ICT as members of the global village.

2. Technology transfer

Technology Transfer (TT) is well-defined in the literature. ATTC (2002) describes TT as *'the process of utilizing technology, expertise, know-how or facilities for a purpose not originally intended by the developing organisation. Technology transfer thus implies that a technology developed for one sector is then used in a totally different area'*. Belonging to the same family as TT, other terms have been popular, such as: Appropriate Technology, Intermediate Technology, Alternative Technology, Sustainable Technology, Capacity Building, etc. These will not be addressed here.

Underlying all of these terminologies is the fact that knowledge and/or skills are being taken from one location to another where such can be used to better the life and living conditions of a group of people or an entire nation. Backer was quoted by ATTC (2002) to have described technology transfer as the transmission of information for the purpose of achieving behaviour change, a description that is similar to that given for learning. The advent of e-learning or Web-learning has thus contributed tremendously in the transfer of skills and knowledge across the global village. The possibilities and effectiveness of e-learning have been instrumental in the improvement of education, training, and manpower development of institutions, organisations at various levels and in different places in developed and less-developed countries. The obvious advantages of e-learning are its "anywhere, anytime" learning possibilities and learner autonomy and flexibilities (Rosenberg 2001).

2.1. Learning, language and e-learning

The human mind is so difficult and complex to understand, that research in this field must be taken as an on-going process which attempts to unravel the mystery. While no psychologist has been able to pinpoint exactly when a learner, in a learning situation, actually commences to attend to the learning stimuli presented, it remains the duty of a teacher to create a conducive atmosphere and environment, and for the stimuli not only to captivate the learner but also to maintain and sustain attention throughout the learning session.

Attention is drawn to two models, which are of vital importance in school learning. The information processing and communication models reveal the attributes of a learner (internal processes) and his/her environment (external factors) and those barriers and impeding forces that can inhibit learning (Gagne et al. 1992).

There is a need, however, to first explain a few terms which relate to teaching and learning. These will, among others, provide answers to the following questions: What is learning? What is e-learning? Who is the e-learner? What is language? What is communication?

3. Learning and instruction

Learning is the essence of education and the literature is replete with definitions, meanings and the complexities of this subject. The *Oxford Advanced Learner's Dictionary* defines learning as *'knowledge that you get from reading and studying'* (Wehmeier 2000). It suffices also to state that learning is a change in disposition which is not attributable to growth or maturity and which can be retained over time. In order for change to occur in the behaviour of a learner, it must be as a result of a stimulus or stimuli received by the individual. The impact of the stimuli must be meaningful, strong, and capable of making the desired change to occur within a short time. Stimuli come in various forms and configurations. These are usually communicated in any or a combination of the following ways: symbolic, verbal, spoken, written and nonverbal in the form of body language.

Learning is known to result from an interaction of symbolic experiences, which may be verbal or nonverbal cues. These are extensively used as means of communication in education. Learning can also occur through vicarious and direct experiences. It is essential to stress the importance of visual literacy in teaching and learning. Visual literacy is the ability of an individual to accurately interpret visual messages and to apply the messages.

We now have to consider what we mean by e-learning. E-learning refers to the use of Internet technologies to deliver a broad array of solutions that enhance knowledge and performance. E-learning is based on three fundamental criteria:

- e-learning is networked, which allows instant updating, storage, retrieval, distribution and sharing of instructions and/or information;
- e-learning is delivered to the end user via a computer using standard Internet technology;
- e-learning focuses on the broadest view of learning – learning solutions that go beyond the traditional paradigms of training.

(Rosenberg 2001: 28)

The e-learner can then be described, from the foregoing, as the individual who interacts with the resource materials for e-learning with the purpose of acquiring knowledge and information. Let us now consider the language factor in learning.

4. Language factor in learning

Language is a unique and powerful means of communication. It may be verbal or nonverbal and it occurs in both living and non-living things. Inanimate objects do communicate as well. They react and respond to stimuli around them. It is language that makes us human and which enables us to work together in all activities of life. It is a rich, varied, and very flexible instrument, playing a major part in our daily life. Though distances separate individuals, language and communication can bridge the gaps in feelings, thought, intentions and plans, as ideas can easily be articulated. The importance of language is underscored in the statement that "the mastery of the word is

the mastery of the world". Word power is indeed magical and language (verbal and nonverbal) remains the most important factor in communication. Language can make or mar an e-learning system if attention is not given to its form and structure in the design of Web materials.

Kelliny & Kelliny (2002) state that the first challenge in the use of the Internet is to have a working knowledge of the English language, as English is the universal language of communication, information technology (IT), business, technology, industry, advertising, entertainment and leisure.

5. Communication in learning

Communication is the activity or process of expressing ideas and feelings or of giving people information. Effective communication is the key to reaching the highest level of learning. This is particularly important in e-learning where the teacher is 'virtual'. Even in ordinary conversations, when communication breaks down, we feel helpless, frustrated, angry and mystified. No form of transfer is possible without proper communication and understanding.

Communication is a two-way process between the source (teacher / e-teacher) and the destination (learner / e-learner). The process is not complete without the inclusion of the message, channel of information flow, feedback and possible sources of noise, which may render communication ineffective. Between the source and the destination are the critical activities of coding, encoding and decoding. The teacher must be able to code and encode information, coded information must be relayed to the learner and the learner has to be able to decode information so transmitted. As a complement to this communication model, there is the information-processing model (Gagne et al. 1992), which explains what is involved, from the moment the information is received from the learning environment by the learner, to the time when the response generated is given back to the environment in writing, action or both.

In Web-based parlance there are two main types of communication modes. Asynchronous and synchronous communication characterise much of e-learning approaches and there are combinations of these two as well. Asynchronous delivery refers to programmes that are independent of time as in the case of pre-recorded audio, video and CD materials being used in e-learning. In this case anyone can access the programme at any time and as many times as desired; there is no live component and it may not involve strict scheduling. Communication does not take place in real time. A synchronous approach refers to programmes that are time-dependent, where communication takes place in real time as in 'chats' and 'video conferencing'. Delivery is live and if a learner misses it, communication was in vain, except when the programme is repeated or recorded for later viewing. Whichever approach is adopted, the language component remains an important determinant of its ability to enable learning and transfer of skills.

6. Conclusion

Learning, language and communication constitute an intricate Web in education. They are inseparable ingredients in e-learning. Change is a phenomenon that human beings have to deal with in every phase of their lives. The issue of transfer, be it of technology or knowledge, in one form or another, carries with it a change in state, form or shape. Spencer Johnson (2002) has proposed an amazing way to deal with change in business and personal life in *Who Moved My Cheese*? E-learning has come to stay and we, as educators, must learn to move with the "cheese". E-learning is not just the "next big thing" in the learning domain, it has become the "now big thing" in business and in education.

In the area of knowledge, which may be multilingual in nature, the value of language is in knowledge and skills acquisition and transfer is of prime importance especially in the virtual realm. Unlike in the past when technology transfers were more physical and tangible in nature, today, learning and knowledge transfers are achieved via cyberspace (electronically). This bestows on the learners a high degree of autonomy and flexibility, which gives them the freedom to learn anywhere and anytime. It is therefore the quality of the language(s) used that will account for the ultimate success of technology transfer from one part of the world to another.

The language used in the text must be according to the level for which it is prescribed and of appropriate vocabulary level. The use of technical terminologies must be carefully explained and the text density must be such that it will not overload the e-learners. Sentence complexity must match the level of the learners for which material was designed. Concern for language is perhaps most fundamental in a successful e-learning system as well as in all training and educational endeavours. Language factor in e-learning remains the major determinant of whether or not and how far the e-material will be successful. There is need for ongoing research in language appropriateness and simplicity in e-learning materials and development.

References

ATTC. 2002. *Addiction Technology Transfer Center*. <http://www.nattc.org/>

Gagne, R.M., L.J. Briggs and W.W. Wager. 1992. *Principles of Instructional Design*. New York: Harcourt Bruce Jovanovich College Publishers.

Kelliny, W.W.H. and I.M. Kelliny. 2002. Is the Internet a tool for Linguistic and Cultural Dominance, Immersion or Integration? In W.W.H. Kelliny (ed.). *Surveys in Linguistic and Language Teaching III: E-Learning and E-Research. European University Studies. Series XXI Linguistics*. Berlin: Peter Lang.

Rosenberg, M.J. 2001. *e-Learning: Strategies for Delivering Knowledge in the Digital Age*. New York: McGraw-Hill.

Spencer Johnson, M.D. 2002. *Who Moved My Cheese? An Amazing Way to Deal with Change in your Work and your Life*. New York: Putnam.

Wehmeier, S. (ed.). 2000. *Oxford Advanced Learner's Dictionary*. Oxford: OUP.

Collaboration Between PanSALB and Terminology Structures

Mariëtta ALBERTS
Manager: Lexicography and Terminology Development, PanSALB, SA

1. Introduction

The availability of dictionaries is an indication of a country's social, cultural, economic, scientific and technological development. As contact leads to communication, proper intercultural communication is a high priority in a multilingual society. In multilingual societies the provision of monolingual, bilingual and multilingual dictionaries can contribute significantly to the improvement of communication between language groups and to the maintenance of standards in the various languages. The language communities of all the official South African languages need dictionaries, not only to bridge the communication gap between them, but also to document and preserve the rich variety of their languages. Languages can only develop into functional languages in all spheres of life by the development of their terminologies.

Modern lexicography, seen in its broadest sense, is a highly democratic activity. In the past, dictionaries were compiled by individuals or publishers, often with dogmatic views, who selected the words or terms to be included and decided how they should be defined. These prescriptive dictionaries are very different from modern dictionaries, which are based upon empirical data gathered from the various speech communities and compiled by means of collaborative lexicography and terminography.

There are many aspects to co-operative lexicography and terminography. According to Van Schalkwyk (1995) co-operative lexicography in its broadest sense refers to cooperation between lexicographers and other interested parties in all fields of lexicography. This paper only concentrates on the holistic collaboration between the *Pan South African Language Board*'s advisory structures, the *National Lexicography Units* and the *National Terminology Office*.

2. PanSALB's advisory structures

The role of language in redressing historical imbalances in South Africa cannot be overemphasised. In order for languages to assume the role of 'official languages' they should be developed. These languages should have the capacity to operate in various domains: economic, judiciary, education, science and technology. The *Pan South African Language Board* (PanSALB) is required by the South African Constitution to develop the official languages and create conditions for use of these languages (Marivate 2001: 5).

PanSALB is a constitutionally mandated Statutory Body established to oversee the process of language development and the protection of language rights. PanSALB was established in 1995 in terms of the PanSALB Act 59 of 1995, amended in 1999. The Board was established to promote multilingualism and to develop the official South African languages, including the Khoe, Nama and San languages, and the South African Sign language. The Board operates under three clusters:

- Lexicography and Terminology Development;
- Development of Languages, Especially those Previously Marginalized;
- Linguistic Human Rights and Advocacy.

PanSALB has made important strides in addressing language developmental problems through the establishment of different advisory bodies, namely the *Provincial Language Committees* (PLCs) and the *National Language Bodies* (NLBs) (Molosankwe 2001: 1). PanSALB also established eleven National *Lexicography Units* (NLUs) for the development of monolingual dictionaries for each of the official languages. These structures will assist the process of language development, especially the nine official African languages (Marivate 2001: 5).

The PanSALB structures will assist the Board in achieving its mandate that is:

- to promote multilingualism;
- to develop languages, and
- to protect language rights.

2.1. The Provincial Language Committees (PLCs)

Nine Provincial Language Committees have been established – one in each province. A PLC is a provincial structure with the aim of taking care of the languages of that province. Each PLC serves the linguistic needs of the people by determining the needs of the various local speech communities. It ensures language policy implementation and practice in order to give the necessary advice to PanSALB and to the Member of the Executive Council (MEC) responsible for languages in the province.

A PLC consists of thirteen representatives proportionally representing each language in the province including Sign, Heritage and possibly Khoe and San Languages.

2.2. The National Language Bodies (NLBs)

The National Language Bodies are responsible for providing advice to PanSALB on matters affecting a particular language. There are thirteen NLBs, eleven for the official languages, one for Khoe, Nama and San, and one for the South African Sign Language. An NLB for the Heritage Languages (i.e. Dutch, French, German, Greek, Gujerati, Hindi, Italian, Portuguese, Tamil, Urdu, etc.) will be established soon.

Each of the thirteen NLBs consists of thirteen members from across the country representing the speech community for the specific language group. The NLBs advise PanSALB on issues relating to:

- the development, promotion and maintenance of its particular language;
- spelling, orthography and language standards;
- terminology development and dictionaries;
- literature.

3. The National Lexicography Units (NLUs)

The history of the establishment of the NLUs starts with the introduction of eleven official languages in the Constitution of the Republic of South Africa. The Government supports the development and preservation of languages whether it is within a bilingual policy or a multilingual policy. In the previous dispensation the country had a bilingual policy and the Government supported two dictionary offices, the *Bureau of the Woordeboek van die Afrikaanse Taal* (WAT) in Stellenbosch, and the *Dictionary of South African English* (DSAE) in Grahamstown. With eleven official languages eleven national dictionary offices need Government support.

The eleven NLUs were established according to the revised PanSALB Act of 1999. The NLUs are situated at tertiary institutions within the boundaries of the geolinguistic area where most first-language speakers of the particular languages are living. The two dictionary offices that existed under the bilingual policy (WAT and DSAE) remain where they are.

A Board of Directors consisting of stakeholders was appointed for each of the NLUs. Each of the Boards of Directors employs staff and governs its own NLU that was established as a Section 21 Company. Each Unit has to document, preserve and develop its specific language by compiling a monolingual dictionary and other projects that will assist development. PanSALB finances the NLUs on a monthly basis.

Staff of the newly established NLUs received various forms of training from PanSALB, e.g. training in the principles and practice governing lexicography, and training in various computerisation aspects such as corpus building, the scanning of documents, the analysis of material, including frequency counts, etc. They also received advice on the purchasing of hardware, software, licenses needed, etc. The NLUs receive regular assistance in aspects such as project management, administrative issues, fundraising, marketing, etc.

4. Relationship between the PLCs, the NLBs and the NLUs and the challenges facing these structures

In each province there is an NLB or an NLU, or both, for the official languages with the majority of the speakers residing in that province (e.g. the isiZulu NLB and NLU (ISS) are both situated in KwaZulu-Natal since the majority of isiZulu speakers reside in that province). The PLCs should keep contact with these bodies in order to make them aware of all the language needs that impact negatively or positively in terms of language policy, practice and implementation and the promotion of multilingualism in the province.

Since multilingualism will impact on education, translation and interpreting, and literature, the link between these three types of bodies is inevitable, defining language needs that affect language policy.

The PLCs, NLBs and NLUs play an indirect role of Public Relations in the various provinces on behalf of PanSALB. It is therefore important that they make themselves known to the Provincial Legislatures, Government Departments, NGOs and CBOs.

PanSALB encourages the various structures to form information sharing or publicity partnerships and to collaborate with bodies pursuing similar objectives.

5. Collaboration

5.1. National collaboration

The South African playing field for lexicography and terminology has been defined in a document of collaboration between the *Department of Arts and Culture* (DAC) and PanSALB. The PanSALB Act of 1995 gives a clear picture of the responsibilities of PanSALB. At present PanSALB has three structures that have an impact on lexicography and terminology practices. The eleven National Lexicography Units (NLUs) compile general, monolingual dictionaries for each of the eleven official languages. The thirteen (soon fourteen) National Language Bodies (NLBs) take care of the spelling and orthography and their role in as far as terminology is concerned is to authorize and authenticate the terminology that has been developed by the *Terminology Coordination Section* (TCS), *National Language Service* (NLS), and other collaborating groups. In fact members of NLBs form the basis of the collaborating groups in the provinces in respect of the work initiatives of TCS. The nine Provincial Language Committees (PLCs) take care of the languages spoken in their particular provinces.

In broad terms, PanSALB's NLUs deal with general dictionaries whilst the TCS deals with the professional terminology management principles. PanSALB's NLBs are the legal authority on all terminology developed in South Africa. Their task, as has been mentioned, is to assist the TCS together with subject and domain specialists, in coining and standardisation of term equivalents. Currently there is a harmonious collaboration and cooperation among all these parties. The TCS cannot develop terminologies for the various languages if they do not consult with the NLBs in respect of spelling and orthography, as the NLBs are the authorising agencies. The NLBs have to determine and document the word-forming principles of each of the official languages. Term creation depends on these word-forming principles.

The Terminology Coordination Section (TCS) of the National Language Service (NLS), Department of Arts and Culture (DAC), promotes the linguistic empowerment of all South Africans through terminological contributions that facilitate communication at different levels in various subject fields and domains of activity. The main objective of the TCS is to coordinate the production of terminologies and external

terminological contributions, forge partnerships with collaborators and stakeholders, and to eventually disseminate term lists to users, clients and collaborators.

The terms documented by TCS will be disseminated to the PLCs, NLBs and NLUs, and all terms created by these bodies should be supplied to TCS for incorporation in the National Term Bank. The terminology products developed by TCS will not only be disseminated to these PanSALB structures, but will also be distributed to other government departments at local, provincial and national level. Such distribution will be meant for optimal utilisation of these terminologies. Feedback and comments on the suitability of the terminology will be sought from these organisations and institutions. Only proper dissemination of lexicographical and terminographical information will allow standardised communication in South Africa. The official South African languages can then become functional languages in all spheres of life.

5.2. International collaboration

The South African terminology and lexicography practice can gain much from collaborating with relevant stakeholders via InfoTerm and TermNet. The international experts are willing to share their knowledge and expertise with South African colleagues. South African lexicographers and terminographers can gain a lot through attendance of meetings, seminars, conferences or contact via e-mail or personal liaison with international experts. The new knowledge can then be utilised to the advantage of the lexicography and terminology practices in South Africa.

6. Conclusion

The PanSALB structures are in place. There is willingness for collaboration amongst the various stakeholders who are members of these structures. There is a working relationship between PanSALB, the NLS and their structures. Indeed, there is a perfect workable situation for the development of terminology in South Africa. However, speakers of the official South African and other languages in South Africa will, according to Marivate (2001: 5) play the bigger role as: *'They should take pride in their languages, use the languages in various domains, and exploit the indigenous knowledge systems embedded in these languages in order to avoid cultural stagnation.'*

References

Marivate, C.N. 2001. A word from the CEO. *PanSALB News*. July – September 2001: 5.

Molosankwe, I. 2001. PanSALB's Advisory Structures (PLCs and NLBs). *PanSALB News*. July – September 2001: 1-5.

Van Schalkwyk, D.J. 1995. Co-operative Lexicography. In J. Cajot et al. *Lingua Theodisca: Beiträge zur Sprach- und Literaturwissenschaft (Jan Goossens zum 65 Geburtstag)* (Niederlande-Studien 16/1): 575-586. Münster: Lit.

Terminology Evaluation

Bassey E. ANTIA & André CLAS
Department of Languages and Linguistics, University of Maiduguri, Nigeria & Département de linguistique et de traduction, Université de Montréal, Québec

Abstract: This discussion takes the view that the quality of a terminology product lies in the quality of decision-making in the course of product development. Against this background, we identify seven nodes in the terminology production cycle, then discuss the challenges associated with each node as well as the relative strengths of methodological options. In doing this, we seek to:

- draw attention to some of the issues that call for conscious and informed decision-making in a terminology development project;
- provide a checklist for terminology project managers; and
- influence in some measure the design and conclusions of product evaluations.

1. Introduction

In many contexts, evaluation is all too often seen as a post-production issue. If we rather saw evaluation as rationalisation of decision-making, involving comparing and choosing from among alternative courses of action and alternative effects, it would be clear that evaluation is a recurrent decimal in any kind of production.

In large-scale language planning contexts, it is possible for a terminological product to be doomed even before it ever has to be marketed, precisely because the processes of its conceptualisation and development were not underpinned by a measure of rigorous evaluation – in the sense of a consideration of as wide a range of alternative decision-making paths as possible.

This brief discussion identifies seven decision-making points, then presents challenges and options associated with each.

2. Dimensions of evaluation

Seven evaluation dimensions are identified and discussed with an eye to one or other stage of a terminology development cycle.

2.1. The organisational-personnel dimension

This dimension invites an examination of various terminology development settings and their respective implications in the areas of personnel distribution, project cost, etc. Nykänen (1993) presents a number of personnel-organisational models for a terminology project, using data from the *Finnish Centre for Technical Terminology*. Figure 1 presents Nykänen's five personnel configurations, classified into broad categories (subject-specialist-centred and terminologist-centred).

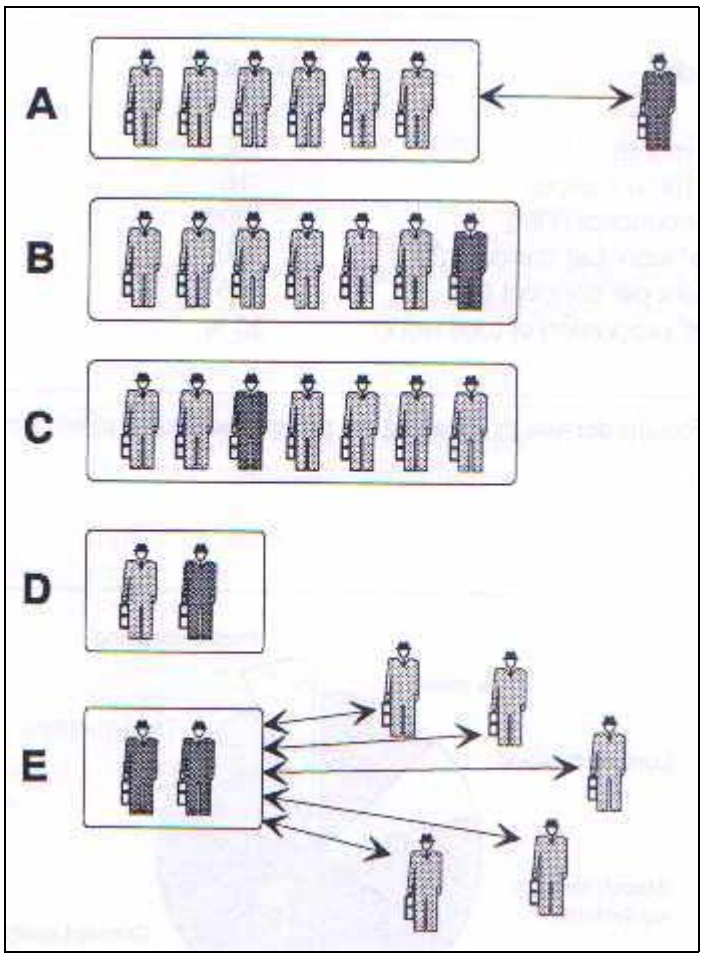

Figure 1: Organisational models (Nykänen 1993)

A. Terminologist as a consultant outside the working group;
B. Terminologist as a member of the group;
C. Terminologist as the secretary of the project;
D. Terminologist producing the vocabulary together with a specialist; and
E. Terminologists working as a team with specialists as consultants.
(Terminologists are the ones with a dark suit.)

Nykänen (1993) compares the time and cost implications of these two sets of organisational models, and obtains the following results:

Table 1: Comparison of specialist- and terminologist-centred models

	Subject-specialist centred models (e.g. B and C)	Terminologist-centred models (D and E)
Concepts per month	6	11
Meetings per 100 concepts	16	6
Total cost per concept (in Finnish Marks)	2300	1200
Terminologists' work per concept (in hours)	5	3.4
Specialists' work per concept (in hours)	4.5	1.3
Terminologists' proportion of total work	50%	70%

It is obvious that an organisational model that places the terminologist at the centre-stage has a quicker project turnaround. While quality was not specifically studied, it is Nykänen's inference from other analyses that *'the quality of terminologist-centred work is the same or even better than the quality of traditional committee work'*.

It might be said that while the product of a specialist-centred model may be adequate for the needs of specialists, a resource created in terminologist-driven settings is likely to be more utilitarian. Subject experts have a knowledge base that helps them bridge gaps in coherence (e.g. in definitions), or that makes it possible for them to take certain kinds of information for granted (e.g. synonyms, collocations, etc.). Such knowledge base may precisely limit the usefulness of such products for outsiders (translators, beginning students of the discipline, etc).

2.2. Linguistic inventory dimension

This dimension invites, among others, a consideration of strategies for determining what words are needed. A basic contrast here is between introspective / intuitive strategies and evidence-based approaches. There is a telling example of the relative strengths of these two strategies. In the West African country of Mali, participants in a Botany course, who were expected to create terms in this area in the local Bambara language, made interesting judgements as to what would be needed. As reported in Antia (2000):

> Most course participants were convinced that beyond equivalents for a few terms like root, stem, flower, leaf, Bambara had no botanical terms. They had no doubt that borrowing [from French] would have to be resorted to. In their minds' eyes, they had already seen *koroli*, *sepali*, *etamini*, *overi*, etc. But on visiting Bambara villages, the team realised the villagers had terms for petals (*feere kala*), corolla (*julakôbô*), pistil (*denkala*), and for ovary (*denso*). (Antia 2000: 35)

This example underscores the importance of rationalising choice strategies for carrying out needs analyses. Even where there are no written sources, focus group discussions can be usefully employed. The use of full texts, indexes of handbooks and textbooks are another source of term candidates. Frequency analyses of indexes of books in related fields of science are being used as the basis for multilingual dictionaries of interdisciplinary science with which Clas is associated (cf. e.g. Kucera et al. 1961, 2002).

The methodology of the ongoing *Canadian Bilingual Dictionary* involves the creation of a corpus, representative of authentic writing in Canadian French and English. A frequency-sorted list derived from this corpus is used to probe existing dictionaries of both languages outside of Canada as a means of identifying what appears to be properly Canadian. In this manner, the dictionary-makers have a basis to decide what to include and what to leave out.

2.3. The epistemic dimension

This dimension deals with the knowledge structure into which the terms being proposed may play any of a variety of roles: fill lacunae, resolve ambiguities, reconcile competing knowledge systems, reflect advances in knowledge, reflect a restructuring of knowledge, and so on. Rationalisation is required in resolving a possible conflict between the old and the new, or on the degree of accommodation of two contending options: novelty (new words, new senses) versus realignments of existing words and senses. In a recent study (Antia et al. *in press*), we compared Fulfulde language terms for cattle diseases as employed by French-trained Fulfulde-speaking veterinary personnel and as used by their clients, cattle farmers (equally Fulfulde-speaking). The acceptation of commonly mentioned Fulfulde terms (as evidenced by elicited descriptions of relevant phenomena) betrays fundamental differences between emic-traditional and cosmopolitan-veterinary knowledge universes. From a communication management perspective, the determination to be made from this data would first be that of a need for resemanticisation (and harmonisation) of existing terminology before it becomes, if at all, one of creating or finding new words.

One consequence of any re-negotiation of existing words and senses is the need for term records to include definitions, and for these definitions to have explicit instructional frames, e.g. 'by X [in this field] is meant PROPOSITION' (cf. Antia 2001).

2.4. The social dimension

Among other pertinent questions here, there is the one on defining the constituency of users of the terminology. There is also the issue of the composition of this constituency. Is the constituency homogenous or heterogeneous? What are the implications for the shape of the terminology to be created?

Two complementary frameworks for addressing some of the above concerns are provided by Kuhn's (1989) analysis of the sociology of dictionary use and Hoffmann's (1985) classification of languages for special purposes. As Table 2 shows, Hoffman categorizes LSPs into five levels of abstraction, and to each level is assigned a set of linguistic resources, context, and communication parties.

Table 2: English adaptation of Hoffman's (1985) classification of LSPs

	Abstraction level	**Resources used**	**Context**	**Communication parties**
A	Highest	Formal language for entities and relations	Fundamental theoretical research	Researcher and researcher
B	Very high	Formal language for entities	Experimental research	Researcher (technologist) and researcher

		Natural language for relations		(technologist) or / and technical assistant
C	High	Natural language Very high terminology content Controlled syntax	Applied research and technology	Researcher (technologist) and technical / scientific head of production
D	Low	Natural language High terminology content Relatively uncontrolled syntax	Material production	Technical / scientific head of production and skilled worker and foreman
E	Very low	Natural language Some terms Uncontrolled syntax	Distribution	Representative of production and sales representative and consumer

Defining the constituency from the above or some other classification provides some guide as to what terms to include or not to include. Note that terms may derive from natural language just as they may come from formal language (e.g. ∑, ≈, 0).

2.5. The textual-discourse dimension

The textual-discourse dimension is a functional criterion or (in a narrow sense) an ergonomic issue. Ergonomics have to do with human-factors engineering aimed at efficiency of interaction between design and user. In this sense, the textual-discourse dimension invites a consideration of options for determining or predicting usability of the terminology from a number of user-perspectives. The contrast is basically between out-of-context determinations or predictions and in-context judgements.

It is common for evaluations to be elicited on the basis of a list of discrete terms, with no context. It is also common for evaluators to use the criterion of knowledge (as may be tested on direct questioning or through a request for (back)translation) as basis to comment on the functionality of the term. In a study by Askira (1995), a low assessment of a term by respondents elicited (from the translator who created the term) a reaction that suggested he, the translator, was concerned about how whatever term was proposed fitted into the surrounding co-text. It is reflective of this neglect of textual-discourse dimensions that in several knowledge, attitude and use studies, claimed usage is consistently lower than claimed knowledge.

In the legislative terminology project reported in Antia (2000), the textual-discourse dimension is investigated and addressed at a number of stages of project development. An existing terminology resource is given to the study's subjects who have to use it to translate a text on legislative matters. They verbalise their thoughts in the translation process. Analysed transcripts reveal problems associated with the

use of terms. The creation of a new terminology is therefore guided by these observed problems, among other issues.

2.6. Psychological dimension

This dimension deals with frames of intellectual and affective cognition, this being an issue in concerns of information processing (such as pattern recognition, mnemonics, etc.), and identification (with the novelty proposed, that is, willingness to accept novelty as useful and/or desirable for a particular function). It is often the case that public expectations are contradictory. A particularly notorious spectrum is the one with identity and intelligibility at opposite ends. Everyone wants a dictionary that is at once a reflection of what is peculiar to a local environment and general. As attempts to document the French spoken in Quebec show, two dictionaries with fairly different outlooks have elicited scathing criticisms from the same public (Clas 2001). The *Dictionnaire du français Plus* elects to be descriptive, documenting French as it is used in Quebec. Its selection of entries is also quite middle-of-the-road on the Quebec-France stretch, and as such it is able to maintain a degree of cross Atlantic intelligibility. On the other hand, the *Dictionnaire du québécois d'aujourd'hui* is heavily tilted towards North American realities or French Canadianisms, and it does not make the distinctions which a descriptive dictionary in France might make between language levels.

Another way in which the psychological dimension needs careful thinking through is in the area of principles of term formation. When the latter involves borrowing of lexemes or structure (calques), guidelines may be derived from analysing contemporary attitudes to race, ethnic or communal relations as documented in public discourse. Sensitivity to these relations should not of course translate into a hostage situation. After listening to a talk on the use of term profiles in reference languages for secondary term formation (cf. Antia 1995), a South African language planning official regretted that term formation models in Afrikaans, Dutch, etc. were unlikely to be acceptable to speakers of African languages in that country. The doubt at the time was obviously a consequence of that country's experience in which language has historically been a theatre of socio-political conflict.

2.7. The public relations dimension

The need to foster dialogue between the project management team and the public targeted by the product is perhaps underscored by the weak quality of argumentation associated with dismissive reviews of neologisms (e.g. such as created by France's Ministerial Terminology Commissions), and the scathing criticism which two differently conceived dictionaries of Canadian French elicited from the same audience (cf. § 2.6). Depecker (2001) observes that criticisms of neologisms in the French press

are occasionally unsubstantiated. Where they are, they are based on personal preferences couched in irony.

Decisions on dialogue with the public are called for in marketing the product idea, the processes of its production and the end result. Indeed, each of the other six evaluation dimensions (at the respective stages of application) can constitute a talking or marketing point. Until recently, any project that sought to create African-language terminology for, say, physics had to be preceded by a public relations campaign, such as that embarked on by Cheikh Anta Diop in the 1950s in Senegal (cf. Antia 2000: 29ff). By the time the Government of Senegal promulgated a policy of mother-tongue education, there was already evidence to convince cynics that development of materials in local languages was indeed possible.

3. Conclusion

Although in no way exhaustive, the seven evaluation dimensions illustrated in this discussion give insight into the diversity of terminological needs, cost-effective methodologies, the profile of would-be users, the distribution (as between verbal and nonverbal form) of terms, the format of elaboration (e.g. definitions), the kinds of awareness activities required to elicit positive attitudes towards the project, and so on. The beneficial use to which this information can be put underscores the point that the quality of a product lies in the quality of decision-making in the development stages.

References

Antia, B.E. 1995. Comparative term records: Implications for decision-making in secondary term formation. In G. Budin (ed.). *Multilingualism in Specialist Communication*: 933-963. Vienna: TermNet.

Antia, B.E. 2000. *Terminology and Language Planning: an alternative framework of discourse and practice*. Amsterdam: John Benjamins.

Antia, B.E. 2001. Quality and Competence in the Translation of Specialised Texts: Investigating the Role of Terminology Resources. *Quaderns. Revisita de Traducció* 6: 16-21. (Special dossier on Empirical and Experimental Translation Research.)

Antia, B.E., Y. Mohammadou and T. Tamdjo. (in press). Terminologie, sécurité alimentaire et santé publique. *META: journal des traducteurs / translators' journal*.

Askira, M.G. 1995. *A Linguistic Analysis of the Translation of English into Kanuri and Hausa by the Borno State Electronic Media*. Unpublished M.A. Dissertation, University of Maiduguri, Nigeria.

Clas, A. 2001. L'éloge de la variation: quelques facettes du français au Québec. *Revue belge de philologie et d'histoire* 79: 847-859 (fasc. 3: langues et littératures modernes).

Depecker, L. 2001. *L'invention de la langue. Le choix des mots nouveaux*. Paris: Armand Colin / Larousse.

Hoffman, L. 1985. *Kommunikationsmittel Fachsprache*. Tübingen: Gunter Narr.

Kucera, A. and A. Clas. 2002. Dictionnaire de chimie / Wörterbuch der Chemie, allemand-français / français-allemand. Wiesbaden: Brandstetter.

Kucera, A., A. Clas and J. Baudot. 1961. Dictionnaire compact des sciences et des techniques / Kompakt Wörterbuch der Naturwissenschaften und der Technik, Deutsch-Französisch / allemand-français. Vol. 1. Wiesbaden: Brandstetter.

Kühn, P. 1989. Typologie der Wörterbücher nach Benutzungsmöglichkeiten. In F.J. Hausmann et al. (eds.). *Wörterbücher: Ein internationales Handbuch zur Lexikographie*: 111-127. Berlin: Walter de Gruyter.

Nykänen, O. 1993. Cost analysis of terminology projects. *TermNet News* 42/43: 20-23.

Facilitating Equitable Access to Government Services through Telephone Interpreting

Anne-Marie BEUKES
National Language Service, Department of Arts and Culture, SA

1. Introduction

South Africa is home to a great variety of different cultures and languages. It is estimated that some twenty-five languages are spoken in the country of which eleven have been granted official status in terms of the language provisions of the Constitution. These eleven languages are spoken as home languages by about 98% of the population. The diverse population lives in a geographical area covering about 1.2 million square kilometres.

One of the characteristic features of the South African language landscape is the phenomenon of linguistic disempowerment and domination. Owing to the past policy of official bilingualism there was an unequal relationship between English and Afrikaans (the only former official languages) and the African languages. This gave rise to language domination and a situation that is referred to as "language disadvantage" in respect of speakers of the African languages. This disadvantage had far-reaching prejudicial effects for many of these speakers in terms of their communication with the government apparatus, access to government services, the administration of justice, education and job opportunities.

These factors, i.e. the complex linguistic profile of the country, the size of the country and language domination, gave rise to, among many other things, numerous disparities and inequalities among language speakers and communities in accessing services and information.

2. Telephone interpreting

In a multilingual environment telephone interpreting offers government a cost-effective mechanism to bridge language barriers and to ensure equity in service delivery. This mode of interpreting offers a relatively simple way of accessing an interpreter over the phone in a few seconds.

The advantages are that an interpreter in the language of choice is always available, no pre-booking is needed, there is no minimum call length and geographical distance is effectively eliminated. Telephone interpreting is therefore particularly suited to the complex multilingual South African environment where language facilitation services are required at short notice in emergency situations and at customer service points where the languages needed are relatively unpredictable and the duration of the consultation unknown or even short.

3. The telephone interpreting project for South Africa (TISSA)

3.1. Objectives of TISSA

TISSA is a Cabinet approved pilot project to test the feasibility of introducing such a service to South Africa. The project is driven by the *Department of Arts and Culture* (DAC), in close collaboration with the *Pan South African Language Board* (PanSALB). The objectives of the pilot project are to:

- test the assumptions underlying the design of the service;
- determine the scope of the service;
- establish the needs and costs of the service; and
- gain experience in the development and running of a telephone interpreting service in South Africa.

3.2. Benefits of TISSA

In addition to affirming government's commitment to the language provisions of the Constitution and the Bill of Rights, the benefits of a telephone interpreting service in the South African context are numerous:

- Improved communication with the client.
  Telephone interpreting enhances the ability of government structures to deal with a client in his/her own language and therefore results in improved service delivery.
- Communication capacity of government structures is improved.
  The sustained use of such a facilitatory service over a period of time strengthens the communication capacities of government structures because information is interpreted and explained accurately, service delivery is accelerated and client frustration and waiting time reduced.
- Support for *Batho Pele* initiative.
  It supports the Batho Pele principles of putting people first in service delivery.
- Immediate solution to language problems.
  On average it takes less than a minute to connect an interpreter to participate in a three-way conference call.
- Cost-effective way of managing language facilitation issues in South Africa.
  The estimated cost of providing on-site interpreting in the official languages at all police stations in South Africa during office hours is about R41,600 per site per month as opposed to R800 per site per month for telephone interpreting on a 24-hour basis.
- Easy-to-use system.
  A single call is required from a speaker phone at service delivery points to a call centre in order to connect a three-way conference call involving a government official, a client and an interpreter. The service is available in all the official languages of South Africa on a 24-hour, seven-days a week basis.

3.3. The scope of TISSA

The feasibility of the service at national government level is currently being tested at seventy police stations of the *South African Police Service* (SAPS) and at local government level in some eleven clinics and eight customer service counters of the Tshwane Metropolitan Council.

Four main stakeholders were initially identified for the pilot service, i.e. the *South African Police Service*, the *Department of Land Affairs*, the *Department of Health*, and the *Department of Labour*. The Minister of Arts, Culture, Science and Technology, Dr. B.S. Ngubane, launched the project on 15 March 2002 at the Katlehong police station in Gauteng. Owing to logistical problems the SAPS has been the only national government structure to participate in the pilot project. SAPS is participating in the project because through TISSA the gathering of accurate information and collaboration between the SAPS and the community in fighting crime is enhanced.

An interesting development in SAPS was the expansion of the TISSA service during the *World Summit on Sustained Development* (WSSD) held in Johannesburg from 17 August to 6 September 2002. TISSA was then made available on a 24/7 basis at designated SAPS sites around the WSSD venue for the duration of the Summit. The telephonic interpreting service was available between English and the following languages: Arabic, French, German, Italian, Portuguese, Spanish and Swahili, in addition to the official South African languages.

As regards participation at local government level, the Tshwane Metropolitan Council joined the project in August 2002. The TISSA service has reportedly resulted in an immediate improvement in service delivery to the Tshwane communities.

Owing to a limited budget available for the running of the pilot project, the service was only available during office hours – between 8:00 and 16:00 – in 2002. However, feedback received from police stations indicated that the need for an interpreting service was highest when fewer police officers are on duty, i.e. after hours and over weekends. In September 2002 the Minister of Arts, Culture, Science and Technology approved that the project be extended to a 24-hour, seven-day a week project. The pilot project will be terminated in December 2003. An assessment report regarding the TISSA project will be submitted to the Minister by June 2003. This report will deal with the financial and operational detail and statistics of the project so as to enable him and Cabinet to decide on the viability of establishing a full permanent telephone interpreting service in South Africa.

4. TISSA and the use of technology

4.1. Technology for spoken languages

The technology flow for the TISSA service is illustrated in Figure 1. As regards the user of the service at a government structure the only technology required is an

ordinary "hands-free" speaker phone while the rest of the technology is located at the Call Centre.

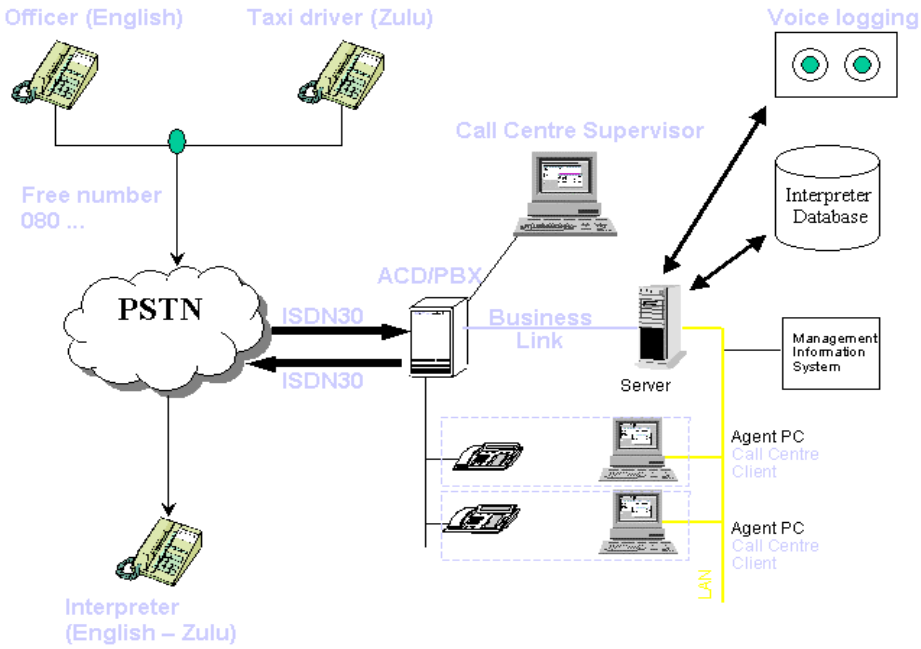

Figure 1: Technology flow in the TISSA project

4.2. Technology for Sign Languages

Telephone interpreting is a well-established concept in Australia and Europe. However, TISSA arguably broke new ground with the introduction of a telephone service that eliminates barriers for users of Sign Languages. A videophone service (cf. Figure 2) for Sign Language interpreting for the deaf was launched by the Minister of Arts, Culture, Science and Technology at the *Bastion Centre for the Deaf* in Cape Town in August 2002.

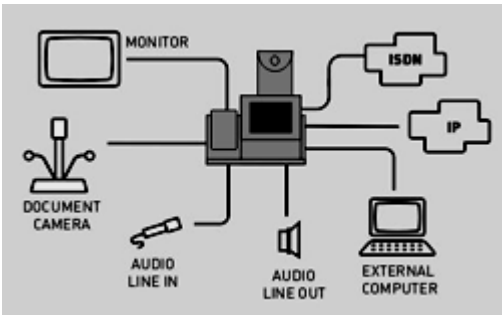

Figure 2: Videophone technology

The videophone is a device transmitting moving pictures and sound to another similar device. It has to be connected to an ISDN line on both sides. The telephone is activated in the same manner as any other telephone – by dialling the number of the other videophone. As soon as the other telephone is answered, both sides are

connected showing the picture on the screen and transmitting the sound. Each phone consists of a camera, a sound card, microphone, speakers and a microcomputer. It is a user-friendly system, as it only requires that an ordinary phone call be made. There is no need to set up a videoconference system.

The TISSA videophone service represents a watershed as far as access to services for the deaf is concerned. Sign Languages are part of a group of languages that were previously marginalized in South Africa. Through the TISSA videophone service deaf people's access to services in their preferred language is facilitated in a unique way.

Videophone technology offers numerous advantages and is therefore a pivotal part of the TISSA pilot service:

- It bridges communication barriers for the deaf by making Sign Language available by enabling a deaf person to communicate face-to-face with a Sign Language interpreter online.
- It eliminates the barriers of physical distance. Access to services is available 'at the touch of a button'.
- The system is portable and only requires an ISDN line. It is also more cost-effective than a videoconference system.

5. TISSA statistics

5.1. TISSA call statistics

After having been in operation for some six months in mid-October 2002 TISSA calls totalled 909.06 minutes (cf. Figure 3). If compared to a well-established service such as the *Australian Telephone Interpreting Service* (with an annual budget of A$ 28 million), which started in 1973 and by 1996 totalled 3,000 minutes per year, the response to the TISSA project is promising.

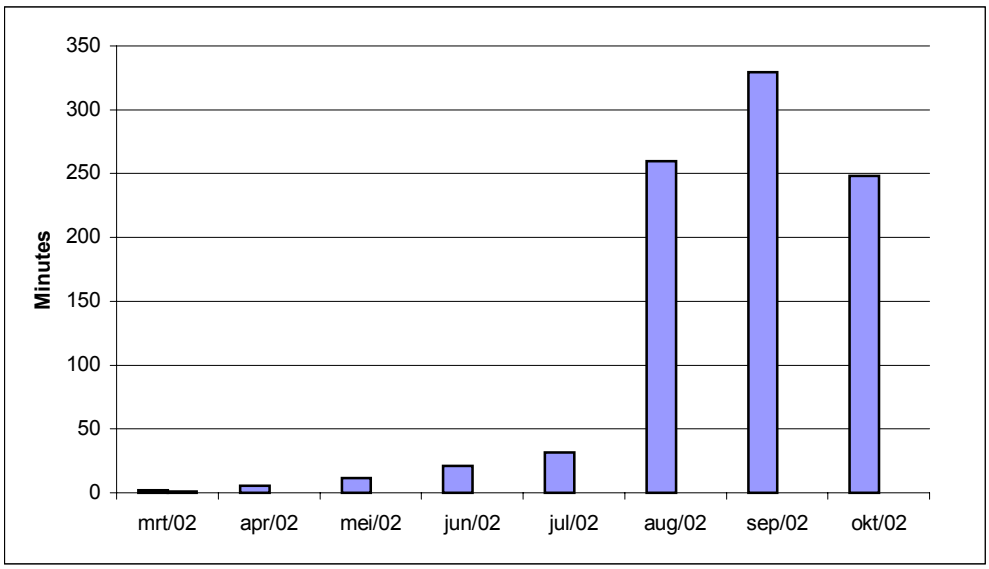

Figure 3: TISSA calls in minutes per month (25 March – 18 October 2002)

5.2. Profile of language use

As regards the distribution of languages in the first phase of the project, Figure 4 clearly indicates that there is a need for all the official languages.

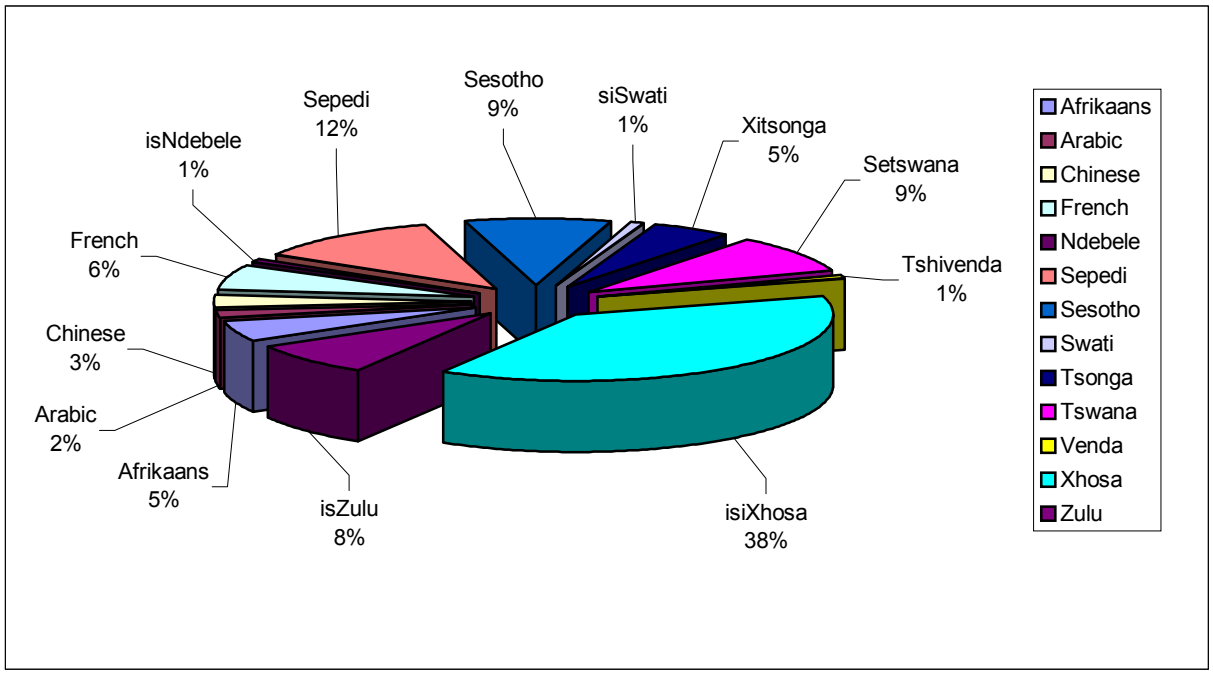

Figure 4: Language profile

From visits to police stations it has however become clear that the attitude of police officers and the location of sites have an impact on the demand for certain languages. It appears that in some police stations officers did not acknowledge their own lack of proficiency in certain languages and therefore did not use the service. This, however, does not seem to be a true reflection of the reality.

The high frequency of isiXhosa is a result of a high level of awareness at the SAPS Knysna, while the demand for Sepedi and other Sotho languages is linked to a high awareness at sites in the Tshwane Municipality.

6. Conclusion

At this juncture in South Africa's quest to transform its language landscape by reversing the legacy of language domination and disadvantage, there is an urgent need to demonstrate feasible ways of making multilingualism work. From the preliminary results of the TISSA pilot project it would appear that telephone interpreting could be a useful mechanism in creating conditions where previously marginalised indigenous languages and also Sign Languages are used alongside each other in all domains of South African life.

Distributed Terminology Management: Modern Technologies in Client/Server Environments

Claudia BLASCHKE
TRADOS GmbH, Stuttgart, Germany

1. Introduction: Local versus distributed scenarios

A Terminology Management System (TMS) is used both as a standalone (desktop PC) application as well as in extremely distributed network environments, depending on user needs. Some functionalities, like searching and editing of terms, must be available at all times to individual users, others, like user and access management, represent requirements of distributed environments.

During the last years TMS became established on the market as an indispensable method of ensuring consistency of mono- and multilingual documentation. Thus high quality product descriptions contribute to overall product quality.

Due to better and better network technology (e.g. flat rate Internet access available for home offices) the call for greater scalability and better performance of TMS without losing proven functionality has become more audible. The following paper will outline what technologies are currently available to meet these requirements.

2. Basic requirements for distributed terminology management

Primarily, a distributed terminology management has to meet all known requirements which are also valid for non-distributed terminology management. These are, among others, concept-oriented, hierarchical data storage, freely structurable data definition, support for all character sets, customisation according to user preferences (display, filtering), powerful search mechanisms, data exchange (import, export), and so on. Distributed terminology management usually assumes a usage scenario in which several users work simultaneously on one centralised data pool. And this is basically independent of being directly (locally) connected to the "master" termbase or not.

A typical scenario of a virtual terminology or knowledge team who interacts globally could look as follows:

1. The development group of an automotive company is located in Paris. This department accesses the master termbase and retrieves and adds terminology.
2. The technical documentation department is centralised in Munich and the authoring language is mainly German. The documentation team also accesses the termbase and looks up corporate terminology in the source language.
3. The Marketing team is spread around the world and needs to access corporate terminology in various languages for the creation of all kinds of marketing material.

4. For the translation of the documentation, the company mainly works with freelance translators who are located in Europe, the USA, Asia and South Africa. The translators work with Translation Memory systems and connect to the master termbase in their relevant target languages.

Depending on the usage of terminology, one could imagine various access media, like power clients for terminologists who are responsible for adding and editing terms and data maintenance, Web clients for read-only users who look up terminology in all available languages. The research teams may be connected to the termbase through their development environments and the freelance translators through Translation Memory systems which retrieve terminology during the translation process.

The following requirements are crucial:

- *Scalability*: A system that has to serve several hundreds or even thousands of users needs to scale – it has to grow with the needs of an organisation. One can start with a small system and extend it accordingly at a later stage.
- *Strong system performance*: Modular, scalable architectures guarantee best performance – no matter what the system load looks like.
- *Ability to involve inhouse users, filial offices, or partner organisations, as well as external vendors and freelance users*: Establishing a corporate language and guaranteeing its usage is only a quality gain that will result in cost savings when everybody in the supply chain can access the centralised data pool.
- *Easy access to data*: User friendliness as known from plain desktop applications is obviously also a requirement in Client/Server environments.
- *Access via various media*: The data storage and its processing engines should be independent from the kind of device or media the data is being consulted with.
- *Low maintenance costs*: Integration into standard IT infrastructure guarantees low TCO, for instance, by relying on established backup processes or thin client technologies (Web browser, w@p access means, etc.).

3. Multi-tier Client/Server architectures

Degrees of scalability: 2-tier, 3-tier, 4-tier

A Client/Server architecture facilitates a better distribution of process load (CPU load) and process logic between several computers. One of the computers could be responsible and optimised for the display of data, another one for data storage. If such a system grows over the course of time, not all of the components would have to be upgraded at the same time, rather, only the server could be upgraded, for example. According to the scaling capabilities of such a system, there are different levels to be identified. Figure 1 outlines the 4-tier model.

2-tier	data display, input and processing		data storage	
3-tier	data display and input (“presentation layer”)	data processing (“business layer”)		data storage (“data layer”)
4-tier	data display and input	data post processing	data processing	data storage

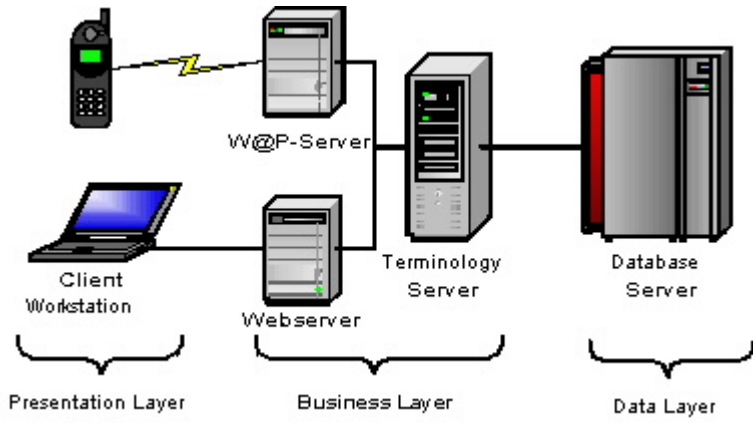

Figure 1: 4-tier model

4. Terminology management in multi-tier model

A multi-tier model may be applied in the following way for processing terminology. Level 1 is responsible for the presentation of the terminology entries, e.g. an Internet browser or even a W@P device. Level 2 is responsible for the data preparation according to the display media, e.g. a Web server. Level 3 is responsible for the actual data handling and logic, e.g. fuzzy searches, filtering, analyses of user rights. Lastly, Level 4 is responsible for reliable and performant data storage, usually taking the form of a database server.

5. Data manipulation based on XML

Particular suitability for processing and storing terminology data

The establishment of XML is a stroke of luck for the machine processing of terminology data. There are several reasons for this:

- *Abstraction of data storage and data display* – one of the premisses of the Client/Server model.
- *Hierarchical structure* – one of the key requirements for terminology management.
- *Support of multilingual data* – a basic requirement for multilingual terminology adminstration.
- *Open and standardised*.

Due to the demand for well-performing and highly reliable systems and their implementation into existing IT infrastructures, data storage must be based on established database systems (Microsoft SQL Server, Oracle 9i or IBM DB2).

However, these systems are not particularly well suited for storing hierarchical data, as their structure is rather table-oriented and following the relational model.

Using XML we can bridge this gulf. So-called XML chunks are stored in the relevant cells of the relational table. Such an XML chunk can be identical with a complete concept or only with a sub branch of a concept, e.g. all parts of one concept belonging to one language or to one term. These rough XML chunks are further processed by using the XML manipulation language XSL (eXtensibel Stylesheet Language) into user-friendly data or documents, e.g. for display in a Web browser.

6. Simplification of data transfer and exchange

The strict pre-determined syntax of XML, as well as the distinction between data and its display, facilitates the implementation of simple interfaces to related systems. The effort needed to exchange and transfer data from one system to another can be estimated easily. This can mean the data exchange from one TMS to another, but also the simple extension of one system in order to support a new requirement like W@P.

7. Support for WAN environments: SOAP

The distributed storage of data is often synonymous to main and satellite data as soon as external users need to have access to the whole or a subset of data of a centralised termbase. Export and import and especially data merging are cost- and timeconsuming and unfortunately computers cannot always fulfil these requirements 100% reliably, like the resolution of homonyms that are being imported. With XML, the new network protocol SOAP ("Simple Object Access Protocol", see http://www.w3c.org) became standardised. It allows access to data sources across networks, for instance, to an Intranet-protected termbase server from a home office via the Internet.

7.1. What is the advantage of SOAP compared to terminology distribution via the Web?

HTML in contrast to SOAP is a document formatting language which allows the publication of terminology but no data processing. A Web browser can be used to retrieve and to display terminology whereas it is not integrateable into translation or text processing applications, e.g. as soon as terminology should be retrieved via Translation Memory systems directly, the transfer of data in form of XML comes into play.

7.2. What is SOAP?

'SOAP Version 1.2 is a lightweight protocol intended for exchanging structured information in a decentralized, distributed environment' (SOAP specification 1.2, http://www.w3c.org). This network protocol is a manufacturer-independent recommendation of the WWW consortium that enables distributed computing on Internet level. SOAP is based on existing Web technologies and infrastructures and

uses mechanisms like HTTP or SMTP for the exchange of data. The exchange format is XML. So-called XML chunks that are transferred as a message construct over a variety of underlying protocols. In the context of a TMS this might mean that a freelance user could have direct real-time access to the main termbase of his customer (within the limits of user rights).

7.3. Usage Scenarios

In advanced terminology, usage scenarios involving various users and user groups, the requirements for a distributed terminology management are data exchange (import and export), merging of entries and data maintenance like duplicates search. Especially the power users of such a system, like the documentation and translation department or freelance translators, may access the termbase through word processing applications, editors or Translation Memory systems. Terminologists may add terminological entries via editors or terminology extraction tools on the fly – no matter where they are.

In the context of worldwide cooperation, asynchronous data is a thing of the past, as there is no need for constant and cost intensive import and export processes of satellite data. Centralised data is always up to date and consistent as error prone replication is avoided. Figure 2 shows a SOAP scenario.

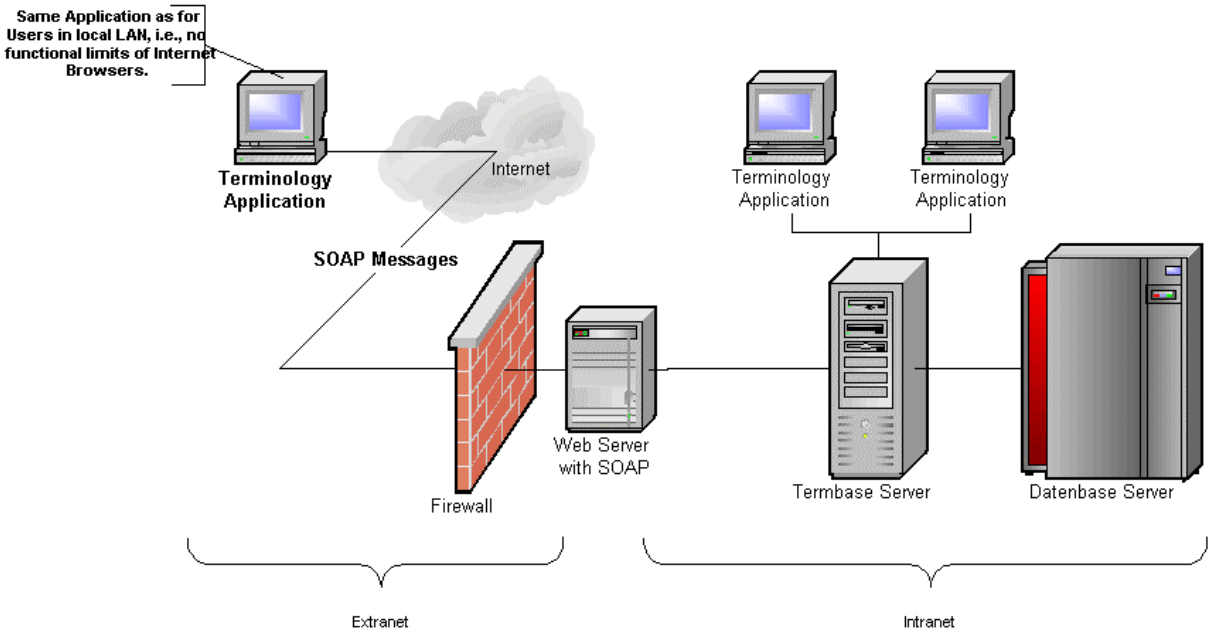

Figure 2: SOAP scenario

8. Outlook

XML-based, scalable TMSs enable the best possible integration into a large number of environments due to their modular and extensible architecture. The more involved users have access to a centralised termbase, the higher the value of a TMS gets, especially when different departments along the overall supply chain, from the production to the translation, have read and write access and changes and updates are

available simultaneously. The integration into knowledge management or CAD systems and therefore the direct integration of all parties involved into the product development and documentation process, as well as the support of media of various kinds, should only be a question of time.

Computational Morphological Analysis as an Aid for Term Extraction

Sonja E. BOSCH[°], Laurette PRETORIUS[‡] & Linda VAN HUYSSTEEN[#]
Department of African Languages, University of South Africa, SA[°#] & Department of Computer Science and Information Systems, University of South Africa, SA[‡]

1. Introduction

The increasing global trend towards automated extraction of terminology from special-language corpora cannot be overlooked in terminography pertaining to morphologically complex languages, such as those belonging to the Bantu language family. The feasibility of automated term extraction in these languages has already been confirmed (Taljard & De Schryver 2002). In this presentation we concentrate on the development of tools which may facilitate automated term extraction in Zulu, and demonstrate their application to a special-language corpus.

2. Aim of the presentation

The aim of this presentation is to explain how a finite-state computational morphological analyser can be used as a tool for the purposes of term extraction from special-language corpora of an agglutinating language such as Zulu.

Alberts (2001: 80) states that terminology *'consists mainly of nouns and, to a lesser extent, verbs'*. The main focus in this presentation will therefore be on nouns as single-word terms, and in the process we pay attention to the following questions:

- What are the challenges posed by term extraction in the languages belonging to the Bantu language family?
- What is finite-state computational morphology?
- How can a morphological analyser serve as an aid for term extraction?

During the workshop a demonstration will be given of a prototype of a morphological analyser for Zulu, based on a sample of an existing electronic corpus *Ingculazi emphakathini*.[1] We shall concentrate on those features that are useful in term extraction.

3. Challenges posed by term extraction in the languages of the Bantu language family

Term extraction in languages belonging to the Bantu language family poses special challenges because of the complex morphological nature of these languages. Compared to a language such as English, for instance, where the variation of word forms is

[1] **Ingculazi emphakathini** 'Aids in the community'. Johannesburg: National AIDS Programme, Department of Health, sponsored by Old Mutual, BP, AIDS Helpline, the European Union, the Open Society foundation for South Africa and UNAIDS.

relatively limited, words in Bantu languages are formed by productive affixations of derivational and inflectional suffixes to roots or stems. In fact the only constant core element in these words is the root.

For example, the following short extracts from *Ingculazi emphakathini* reveal a number of occurrences of the noun *umzimba* 'body', each formed from the basic root *-zimba*, but with different affixes in five of the eight occurrences:

> **Imizimba** yethu inamalunga amaningi, futhi wonke amalunga **womzimba** anemisebenzi ebalulekile ayenzayo. Isibonelo, inhliziyo impompa igazi liye kuwo wonke **umzimba**, ingqondo iyacabanga, amaphaphu aphefumula umoya, amabele enza ubisi, kanjalo kanjalo. Sinesistimu ebaluleke kabi **emizimbeni** yethu ebizwa ngokuthi yisistimu yokuvikela (immune system). Umsebenzi walesistimu wukuvikela **imizimba** yethu emagciwaneni nasezifweni. Futhi yenza ukuba **umzimba** uphole emva kokugula noma ukulimala.
>
> Isistimu yokuvikela ifana namasotsha - ivikela **umzimba**. Ibutho libabuthakathaka lingakwazi ukuvikelana kahle uma amabutho efa.
>
> Igciwane i-HIV liyayibulala kancane kancane isistimu yokuvikela uma like langena **emzimbeni** womuntu.

It should be noted that Zulu follows the convention of a conjunctive writing system. So what we observe in the bold-printed examples, are linguistic words, each consisting of a root surrounded by its affixes. Each word is of a poly-morphematic nature or consists of a number of bound parts or morphemes which can never occur independently as separate words. The two types of morphemes that are generally recognised are **roots** and **affixes**. Roots, which are also called radical morphemes, always form the lexical core (meaning) of a word, while affixes usually add a grammatical meaning or function to the word.

If we therefore wish to extract single-word terms from a corpus for instance, each word in the corpus needs to be morphologically analysed in order to obtain reliable feedback. In this regard Hurskainen (1997: 633) remarks:

> Many tasks in the process of retrieving language-specific information from text can be carried out in more than one way, by using suitable tools nowadays abundantly available. What cannot be substituted by commercial, shareware, or public domain tools is the morphological parser.

The reason is that morphological rules that are applicable to one language, are not necessarily applicable to another. Therefore, the morphological analysis needs to be language-specific. Two problems which are, however, central to any morphological analysis, are the following:

- *morphotactics*: Words are typically composed of minimal meaningful units, namely morphemes which cannot combine at random, but are restricted to certain combinations and orders. A morphological analyser needs to know which combinations of morphemes, or word-formation rules, are valid.
- *morphological alternations*: One and the same morpheme may be realised in different ways depending on the environment in which it occurs. Again, a morphological analyser needs to recognise the correct form of each morpheme.

We shall briefly look at some of the implications of morphotactics and morphological alternations in Zulu, with specific reference to the noun.

Firstly, regarding the *morphotactics* or word formation rules, there are numerous possibilities of morpheme concatenations in Zulu nouns for which possible legal combinations need to be determined. Table 1 gives us an idea of possible morpheme combinations and legal morpheme orders that need to be recognised for the noun category.

Table 1: Morphology of the noun

Other prefixes	Preprefix	Basic prefix	Root	Nominal suffix	Deverbative suffix
ku	u	mu	ntu	kazi	
y	i	si	khathi	ana	
na	i	mi	thwal		o
ka	u		baba		

Roots are preceded by one of approximately thirteen variations of the noun prefix. The noun prefix may then be subdivided into a preprefix and a basic prefix, and may be followed by nominal suffixes such as the diminutive and augmentative, or even deverbative suffixes. Furthermore, the preprefix may be preceded by a number of morphemes such as the copulative morpheme, various adverbial morphemes including the locative, the possessive morpheme, and so forth.

Secondly, let us turn our attention to the rules that determine the form of each morpheme in Zulu, namely *morphological alternations*. Words are mere concatenations of morphemes, but raw concatenation often gives us abstract 'morphophonemic' not-yet-correct words. There are 'alternations' between the raw concatenations and the desired surface word forms.

In the following example, all three words share a common root morpheme, namely *-lomo* 'mouth' which is hardly identifiable in the surface word forms in (1b) and (1c). What makes it difficult to identify the underlying root morpheme on surface level in these examples, is the fact that certain morphophonological processes have taken place.

(1a) *umlomo*
(u-m(u)-lomo)
'mouth'
(1b) *umlonyana*
(u-m(u)-lomo-ana)
'small mouth'
(1c) *emlonyeni*
(e-(u)-m(u)-lomo-ini)
'in the mouth'

A morphological analyser would have to take certain morphophonological processes into account, that is where changes in the sounds of morphemes are based on surrounding phones. It would, for instance, have to be specified in the analyser that palatalisation could occur with certain noun suffixes (diminutives and locatives), as has been illustrated in (1b) and (1c) respectively; and that vowel elision occurs when the basic prefix *-mu-* is followed by a polysyllabic noun root as in (1a), (1b) and (1c), or when the locative prefix *e-* is prefixed to a noun, as illustrated in (1c).

In the following section it will be shown that these challenges which have a bearing on automated term extraction can be addressed within the framework of finite-state computational morphological analysis.

4. Finite-state computational morphology

The Xerox finite-state tools (Beesley & Karttunen 2003) constitute a collection of powerful, sophisticated, state-of-the-art 'programming languages' and algorithms for developing finite-state solutions to a variety of problems in natural language processing, including morphology, phonology, tokenisation, and shallow parsing. The application of these tools to morphology, in particular, is based on the fundamental insight that we can model the complexities of word-formation rules and morphophonological alternations by means of finite-state methods. The Xerox software tools that we use to build a morphological analyser for the Zulu language, namely **lexc** and **xfst**, are briefly discussed.

The purpose of the **lexc** tool, for {lex}icon {c}ompiler, is to specify the required and essential natural-language lexicon, as well as the morphotactic structure of the words in this lexicon. The finite-state network, generated by **lexc**, produces morphotactically well-formed, but rather abstract morphophonemic or lexical strings.

The purpose of the **xfst** tool is to define the alternation rules required to map the abstract lexical strings into correctly spelled surface strings of natural language by using regular expressions. These regular expressions are subsequently compiled into a finite-state network.

Finally, the **lexc** and **xfst** finite-state networks are compiled (composed) into a single network, referred to as a lexical transducer. This lexical transducer represents

all the morphological information about the language being analysed, including derivation, inflection, alternation, compounding and so forth, and constitutes our computational morphological analyser.

Schematically the application of a morphological analyser can be represented as shown in Figure 1.

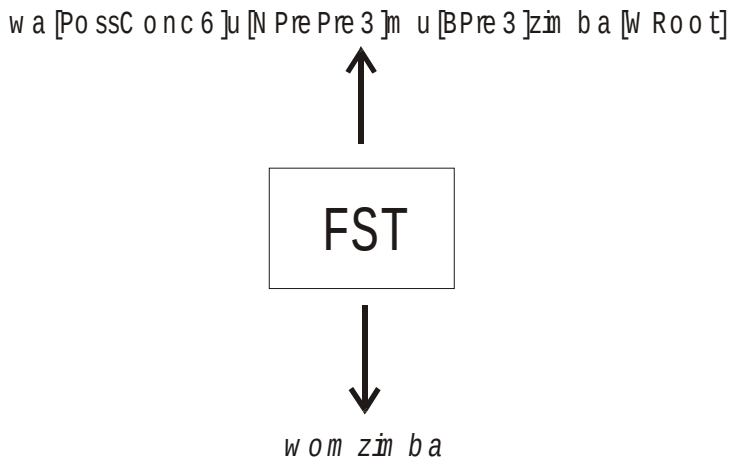

Figure 1: Schematic representation of the morphological analyser

In Figure 1 the morphological analyser maps the morpheme sequence *wa-, u-, -mu-* and *-zimba* to *womzimba* 'of the body'. It should be noted that bidirectionality, that is analysis as well as generation, which is a useful feature of the Xerox tools, is illustrated clearly by the bidirectional arrow. This feature will not, however, be discussed in this paper.

5. How can a morphological analyser serve as an aid for term extraction?

Two of the basic outputs of a specialised electronic corpus that are of interest to the terminographer, are term frequency counts and concordance lines. In a morphologically complex language such as Zulu, in which the only constant core element in words or word forms is the root, as has been illustrated above, it is essential that word roots be identified in order to (i) list the morphological variants of a word form under a single lemma or canonical form, and (ii) list selected canonical forms in a concordance indicating their immediate context to the left and the right. Frequency distributions of word forms and concordance lines are not only useful in identifying single-word term candidates, but also assist in determining parts of potential multiword terms by means of collocation patterns. The extraction of accurate frequency counts and concordance lines from an electronic Zulu corpus both call for each word in the corpus to be morphologically analysed.

A very simple term to extract from the sample corpus would for instance be the noun *ingculazi* 'AIDS', should it appear in its uninflected form. Yet, its frequency would be much higher if all inflected forms of this same noun would also be extracted.

Such inflected forms, however, appear in a word list ranked as words separately from *ingculazi*, each with its own frequency, e.g. *yingculazi*, *zengculazi*, *ngengculazi*, *abanengculazi*, *eyingculazi*, *kwingculazi*, *usunengculazi*, *lengculazi*, and so forth. In fact, all the mentioned words have the same root namely *-ngculazi*.

From our examples it should become clear that in Zulu, scientifically reliable frequency counts and concordance lines are in general difficult to obtain without automated morphological analysis. In the eight words above taken from the sample corpus, the root *-ngculazi* is realised in a different word form each time. In order to identify a root for each of the forms, a process of so-called affix-stripping (Gibbon 2000: 34) is applied by means of the computational morphological analyser.

For the morphological analyser to be able to identify roots, they need to be present in the analyser. Words constructed from roots that are not present, will not be analysed. Referring back to the sample corpus, should the noun root *-ngculazi*, for instance, not be present in the morphological analyser, it would not be possible to identify this root in any of the eight different word forms as indicated above.

In order to systematically update and extend the root list of the morphological analyser, reflecting the dynamic nature of the language as well as the well-known principle that terminology compilation is an on-going and repeated activity (Sager 2001: 763), the Xerox finite-state tools include a useful feature, namely the option of building a so-called guesser. The guesser is a variant of the morphological analyser that contains all phonologically possible roots (XRCE 2002). This guesser variant of the morphological analyser is a particularly useful computational tool for exploring (new) language corpora. By applying the guesser to any corpus as a potential source of new word roots, new (as yet unlisted) word roots are detected, analysed and marked for possible inclusion in the current word root list.

Ultimately it will be the responsibility of the terminographer to determine whether such roots form part of the terminology of the particular subject field, since *'all term-extraction and retrieval systems are still in some sense support systems that offer proposals to the terminologist who must then decide how to proceed, often in consultation with a domain expert'* (Ahmad & Rogers 2001: 740).

6. Conclusion

The complex morphological structure of a language such as Zulu necessitates morphological analysis as an aid for automated term extraction. In order to facilitate reliable and scientific term extraction from a special-language corpus even further, the guesser variant of the morphological analyser is a particularly useful computational tool for identifying possible new terms. Sager (2001: 765) confirms that text analysis of large corpora can be used to identify new terms, and that while this feature of terminology compilation is still in its infancy stages, its significance will increase greatly in the future.

Acknowledgements
The authors would like to acknowledge the financial support of the *National Research Foundation* (NRF) for the project 'Computer Analysis of Zulu', the contribution of *Xerox Research Centre Europe* (XRCE), whose software and training are a vital component of this project; as well as the continued support and expert advice of Dr. Ken Beesley (XRCE).

References

Ahmad, K. and M. Rogers. 2001. Corpus Linguistics and Terminology Extraction. In S.E. Wright and G. Budin (eds.): 725-760.

Alberts, M. 2001. Lexicography versus Terminography. *Lexikos* 11: 71-84.

Beesley, K. and L. Karttunen. 2003. *Finite-State Morphology: Xerox Tools and Techniques*. Stanford: CSLI Publications.

Gibbon, D. 2000. Computational Lexicography. In F. Van Eynde and D. Gibbon (eds.). *Lexicon Development for Speech and Language Processing*: 1-42. Dordrecht: Kluwer Academic Publishers.

Hurskainen, A. 1997. A language sensitive approach to information management and retrieval: the case of Swahili. In R.K. Herbert (ed.). *African Linguistics at the Crossroads: Papers from Kwaluseni*: 629-642. Cologne: Rüdiger Köppe.

Sager, J.C. 2001. Terminology Compilation: Consequences and Aspects of Automation. In S.E. Wright and G. Budin (eds.): 761-771.

Taljard, E. and G.-M. de Schryver. 2002. Semi-Automatic Term Extraction for the African Languages, with special reference to Northern Sotho. *Lexikos* 12: 44-74.

Wright, S.E. and G. Budin (eds.). 2001. *Handbook of Terminology Management*. Amsterdam: John Benjamins.

XRCE. 2002. References to finite-state methods in Natural Language Processing. <http://www.xrce.xerox.com/research/mltt/fst/fsrefs.html> (Last accessed on 04/11/2002)

The Practical Use of Knowledge Management Theory in Terminology Management

Mark D. CHILDRESS
SAP AG, Walldorf, Germany

1. Introduction

Terminology management, like all management approaches, focuses on identifying the critical handoff points where the person-based, technical, and organisational processes do not work as expected. Determining which resources are required to solve certain problems is usually easy – getting and mobilizing those scarce resources is the hard part.

In the ongoing battle for resources, knowledge management (KM) offers useful insights when it comes to prioritising solutions. This paper is a brief summary of how an organisation can benefit from the practical application of KM theory to identify these critical points in its terminology management. The terminology manager will then find it easier to decide where to apply solutions.

2. Knowledge management and terminology management

Knowledge management in the most general sense is an organisational approach to generating value from intellectual and knowledge-based assets. Effective knowledge management supports innovation and optimises processes by explicitly supporting the exchange and comparison of information from related – and unrelated – branches of the same organisation. By nurturing the free flow of knowledge and making knowledge capture and distribution an integrated part of all processes and procedures, an organisation will reap benefits, sometimes unexpected ones, as ideas multiply and employees are rewarded for sharing their own knowledge with others.

Similarly, terminology management on the organisational level is the application of practical tasks carried out to standardise terminological information. Most organisations do not undertake terminology management for its own sake. The work is meant to add value to organisational information and describe shared concepts inside and outside of the organisation. Organisations will tend to standardise terminology, deliberately or unconsciously, in order to better achieve the organisational goals. It is of course far better to manage this activity rather than to let it take its own unplanned course.

Based on these definitions, terminology management can be described as a specialised form of knowledge management. It involves reviewing (capturing) the concepts (knowledge) created by individuals in an organisation and making these concepts available to others (distributing) by means of various delivery channels defined and supported by the organisation.

In turn, knowledge management fosters many different processes of knowledge creation, knowledge capture, and knowledge distribution throughout the organisation. One important aspect thereof is the creation, review, and distribution of knowledge-rich terminology, which serves as the smallest unit of knowledge describing an individual concept. The linkage of concepts that results from an organisation's collected terminology is the foundation upon which more knowledge is created, captured, and distributed to others.

2.1. Terminology creation as knowledge creation

Organisations do not create knowledge; people create knowledge. The organisation provides the environment in which knowledge finds its expression. Organisations can stifle knowledge creation, sometimes unintentionally, or they can intentionally nurture a culture in which it flourishes.

Thus the first step in terminology management is to coax or support the creation of terminology as a form of knowledge useful to others in an organisation. Especially in the early stages a great deal of the organisation's terminology resides only in the heads of the employees. The organisation must ensure that all employees understand what terminology is and why standardised terminology is so important. Moreover, it must support terminologists trained in terminology creation and review in their special subject areas.

In a more practical vein this means:

- Defining what 'terminology' means for the organisation – what is a 'term', what is an 'abbreviation', and so on. The organisation then ensures that these broad definitions are clearly understood and available to all employees.
- Formulating specific terminology standards based on ISO standards for each language required by the organisation. These standards emphasize general standards applying to all languages and include specific exceptions and special standards tailored to the needs of each language.
- Providing terminology information and training based on roles in the organisation. This includes general information for all employees and terminology management training for those working directly with terminology issues.
- Obtaining official support from management for these standards and policies, usually as part of overall quality management.

2.2. Terminology review as knowledge capture

Knowledge capture must cross the greatest mental gap. A veritable canyon lies between what the employee knows and the methods by means of which this knowledge is made understandable and repeatable to the rest of the organisation. Much personal knowledge remains unformulated because organisations directly or indirectly discourage their employees from sharing it, or at least do not actively encourage them.

The crux of terminology management is thus the review process. The organisation must provide systems and processes in order to document, standardise, test, and revise or reject the concepts and the related terms arising from terminology creation.

For the organisation this means:

- Teaching employees methods and theory in order to recognise and test the terminology concepts arising in specific subject areas.
- Providing tools for documenting terminology and ensuring a standardised workflow for review work. In best-case scenarios, such tools check documented terms for compliance with standards in order to save review time and costs.
- Defining or adapting standardised processes in order to ensure the employees have enough time to review terminology in addition to their standard duties.

2.3. Terminology delivery as knowledge distribution

Not only is it impossible for a single employee to personally share knowledge with everyone else in an organisation, even a small one – the employee is even less able to share knowledge on a personal basis with other employees in other organisations. Therefore the organisation must provide formal, operational structures or tools, media, and methods for distributing knowledge.

Individuals may, of course, explain terms to individuals or small groups, but the ultimate goal of any terminology management strategy is to deliver a collection of standardised terminology to predefined user groups or roles.

Generally speaking, internal terminology delivery needs to be as real-time as possible for all employees – the quicker the terminology is standardised, and the quicker this is made available, the better. External terminology delivery may be real-time or somewhat further removed in time, depending on need.

The organisation must also determine access needs. Most roles will be interested in, and limited to, read-only access; others will be involved more or less in terminology creation and review activities. A special internal user group is responsible for the terminology review process. The employees in this group require direct editor access to the terminology itself and additional training in order to make the end terminological product benefit the organisation.

From a delivery-based viewpoint this means:

- Devising a strategy for operational access to and delivery of the terminology based on printed or electronic modes.
- Determining who needs access to the terminology at which stages (creation, review, and final), and to which degree (internal / organisational or external / public).

2.4. Continual process improvement

Knowledge management is a strategy of constant examination and improvement of processes. It is not enough to look at the knowledge processes once, make improvements once, and let it go at that. A great deal of activity remains bound up in monitoring the ever-changing organisational situation, gathering feedback, and striving for improvements. The organisation must provide for constant analysis of problems and for solving them by applying new tools and techniques.

Similarly, terminology management is not a perpetual motion machine with a self-repairing process flow. The knowledge management approach helps identify the factors in terminology management which would benefit the most by concentrating resources on problem solving.

Specifically, this means:

- Servicing and supporting the terminology management processes.
- Instituting a terminology management forum in which problems and issues are dealt with on an operational basis and which is responsible for a general review of all processes on a regular basis.
- Maintaining a steering committee of key managers and coordinators whose purpose is to ensure higher-level management support for special terminology projects and initiatives.

3. Conclusion

No one can avoid or afford to ignore terminology. Terminology is everywhere and the organisation will just have to accept that as a fact. Organisations thus need well-defined processes to deal with terminology issues, and to continuously monitor and improve them when necessary. This is a far better alternative than ad hoc procedures or random acts of terminology work.

Knowledge management is a practical, effective tool for pinpointing problem areas or recurring situations related to knowledge transfer. It pays to apply this viewpoint to existing formal or informal terminology management processes. The terminology manager will find it far easier to prioritise, to research specific problems, and to develop relevant solutions based on the resources at hand. The solution may be a tool development, changed steps in a process, or additional standards and training information made available to the individuals in the organisation. In some cases a solution solves multiple problems at a single stroke.

Neither knowledge management nor terminology management are magical cure-alls. Constant, open communication between all involved parties is essential. Managerial buy-in and active support for problem-solving initiatives are also a deciding factor. But, applied correctly, both forms of management will go a long way towards improving the terminological health – and wealth – of any organisation.

Multilingual Technology and Technology Transfer

Jennifer DECAMP
MITRE Corporation, McLean VA, USA

Abstract: Problems in multilingual technology and technology transfer relevant to African languages include lack of (i) information about available technologies, (ii) means of easily sharing available technologies, and (iii) input to developers, professional organisations, and standards committees on technology needs. This paper discusses these problems and provides suggestions for circumventing and/or addressing them.

1. Information about available technologies

Many language technologies – particularly word processing – have been developed by or for particular university or government offices (e.g. Hausa for the *United States International Broadcasting Bureau*, aka *Voice of America*) but have not been extensively marketed. There are also workarounds that are not always known by the representatives of the company offering the product, some of which are discussed in the last section of this paper. Lack of information about capabilities and lack of access to the software often results in research groups and software companies providing redundant efforts, at a time when there is extensive new development that is needed.

There are many services and products that can be of help. One is a project called the *Foreign Language Resource Center*, which is a website at http://flrc.mitre.org designed for information exchange about language technology. The site is sponsored by the U.S. government, and extensive global participation is encouraged. The site includes a survey form, which can be filled in and updated by the developer or sales organisation. The survey form includes questions about the functionality, quality, and availability of the software. Input is solicited on current and planned products, prototypes, and research.

The site also includes electronic reports, where the user can request all products in the database for a particular language, all products for a particular technology (e.g., optical character recognition), all products for a particular language and technology (e.g. all Hausa optical character recognition), and/or all products available in certain time frames (e.g. new in the coming year). We are working on a capability for the products to be displayed by functional capabilities (e.g. to provide optical character recognition of handwriting rather than just of print), as indicated by the vendor or developer providing the information. A second stage will be to validate, assess, and rate the information, with the information then marked as having been validated. Feedback will be provided to the vendors and developers.

The survey currently includes a capability for the respondent to provide information on assessments, specifically on the metric, measure, score, date of assessment, and who conducted the assessment. This information is provided in the

electronic reports described above, and where possible, the original assessments are provided as links. There is, of course, a cautionary note in the report that information cannot be easily compared across assessments, particularly when assessments are conducted with different methodologies. For selected products that meet a wide range of functional requirements, my company, MITRE, plans to conduct or help structure assessments that can be compared across products, thus enabling better relative ratings.

We welcome any input on the types of data we collect, the languages for which data is collected, and the products. We also welcome any input of data and any use of this resource.

2. Means of easily sharing available technologies

Lack of information about technologies of course, has been one of the key barriers to sharing technologies. However, there often is not the support structure to make computer code more robust, to add documentation, and to provide assistance. Thus people may give their code to friends or colleagues, but are reluctant to make it more broadly available because of the impact to their time and resources. Conversely, users are often reluctant to try programs available on the Internet, particularly if it is untested and/or undocumented.

There is a further problem that products developed for English and Western European languages do not always work well and/or cannot be easily adapted to substantially different languages. Many companies (e.g. TRADOS, Microsoft) are increasingly adopting standards for the *International Organisation of Standards* (ISO) and the *World-Wide Web Consortium* (W3C). However, until 2002 there has been little African input to such products or standards, either in the design or testing stages. Substantial time and work is needed to review standards in light of the requirements of the many African languages to ensure that these standards – particularly ones used in software development – are conducive to meeting the needs for applications in these languages.

Where software is available without licensing costs and would be helpful to this global language community, we would like to make it available on the *Foreign Language Resource Center* site through linking to relevant sites, or through providing the software to be downloaded. Information and/or software can be sent to Jennifer DeCamp at jdecamp@mitre.org. In some cases, with the permission of the developers, we would like to make the freeware more robust by providing documentation and instructions for installation and use.

For the U.S. government, we also provide a help desk, where members of the U.S. government can e-mail or telephone with requests for information or for help with software problems (e.g. incompatibility with other products, inability to perform as advertised, etc.). A similar help desk is provided by the Austrian government to make multilingual products more accessible to businesses. Such a help desk may also be

useful for South Africa, particularly in this current stage of rapid emergence of technological capabilities in African languages.

3. Input to developers, professional organisations, and standards committees on technology needs

In order to have commercial products that support word processing, optical character recognition, searches, terminology management, and other functions in African languages, it is often necessary to provide input to developers, professional organisations, and standards committees on technology needs relevant to these languages. One approach is to participate in organisations that develop standards for industry use. The *Unicode Consortium* and the *Localisation Industry Standards Association* (LISA) are organisations that were formed by industry to develop standards for their use. Meetings of these organisations include product managers and marketing managers from a wide range of companies. Both organisations welcome new members, although there are steep membership fees. There are also open-source software groups that welcome participation of all interested parties, particularly people willing to take on new development projects. Examples include the LINUX I18N (Internationalisation) newsgroup, which develops new software available free to all parties. Another example is Netscape, which has made its source code for the basic browser, Mozilla, available on the Internet for developers to expand upon. A particularly productive area for expansion has been international extensions for word processing and localisation. These organisations are particularly relevant for word processing, browsing, and localisation applications.

In the terminology area, there are more traditional standards organisations such as *ISO Technical Committee 37* and *W3C Internationalisation* (I18N). Such committees have extensive participation by industry and by government organisations. Organisations such as InfoTerm and TermNet include terminology publishers and interact with software developers. Professional organisations such as TAMA and *Terminology Knowledge Engineering* (TKE) provide opportunities for exchange of information between users, publishers, software developers, and standards representatives. LISA is also a good forum.

For machine translation, there are numerous organisations, including the *International Association of Machine Translation* (IAMT). For many other language technologies, a good forum is ELSNET (the European Network of Excellence in Human Language Technologies).

Direct approaches are also effective. Many software companies welcome input on user needs, although a frequent question from such companies concerns the amount of business that could be expected if the requested capability is added. There is also the approach of enlisting broad support for capabilities in certain languages, thus enlarging the market for industry. Many universities, government diplomatic services, government health systems, and libraries throughout the world have needs for abilities

to create, edit, display, print, and search text in at least some African languages. By developing joint business cases, it is often possible to get a sufficient business case to interest software developers. If the market for certain language software is small, it helps to reduce the cost to industry by prototyping or developing or paying the company to develop and integrate certain capabilities.

Many companies also welcome beta testers. If a requested capability is developed, it is a good idea for the requesting organisation to at least review it, or preferably beta test it, before the software is released in order to ensure that the capability functions as envisioned. Considerable input to future software releases can be provided in the beta testing process.

4. Conclusion

Africa and the global community interested in language technologies face a significant endeavour in identifying language requirements (e.g. character sets, keyboard layouts, linguistic constraints, etc.); identifying existing technology that can meet those requirements or can be easily adapted to meet those requirements; providing information and support to users through websites, help desks, and other sources; ensuring that existing and planned technologies provide maximum functionality and easy use with commercially-available products; ensuring that international standards meet the linguistic and cultural requirements of African languages; and addressing gaps. This is not the scope of work for a single person; it is not the scope for a few busy volunteers with other commitments. This work will require commitments of funding, time, and dedication, but should be highly instrumental in advancing technologies and technology use throughout Africa.

Recommended websites

- http://www.dictionary.com & http://www.yourdictionary.com
  Sites with useful electronic dictionaries.
- http://www.multilingualcomputing.com
  A site with extensive information on multilingual technology.
- http://www.systransoft.com & http://www.systranet.com
  Sites with ability to use free rough machine translation, mainly in European languages and English. The second site includes a capability for the users to add their own terms.
- http://www.elsnet.org/list.html
  Site with extensive information on resources (predominantly European) for language technology expertise.
- http://www.unicode.org
  The Unicode site with extensive information on this standard.
- http://www.eamt.org/compendium.html

Machine translation resources, available through the website of the *European Association of Machine Translation* (EAMT).

- http://www.lisa.org
  Site of the *Localisation Industry Standards Association* (LISA), with information on localisation and on terminology standards.
- http://www.sign-lang.uni-hamburg.de/bibweb/Keywords/African-sl.html
  Information on sign languages for African languages.
- http://www.sas.upenn.edu/African_Studies/K-12/menu_EduLANG.html
  Information on African languages.
- http://www.ethnologue.com
  Information on languages, searchable by language name and/or country.
- http://polyglot.lss.wisc.edu/lss/lang/african.html
  Information on African languages.
- http://www.balancingact-africa.com/news/back/balancing-act_69.html
  Article on technologies for African languages.

Towards Strategies for Translating Terminology into all South African Languages: A Corpus-based Approach

Rachélle GAUTON[°], Elsabé TALJARD[‡] & Gilles-Maurice DE SCHRYVER[#]
Department of African Languages, University of Pretoria, SA[°‡#] *&*
Department of African Languages and Cultures, Ghent University, Belgium[#]

1. Introduction

The single biggest problem that translators who translate from a language such as English into the African languages have to contend with is the lack of terminology in the African languages in the majority of specialist subject fields. The relevance of terminology theory and practice for translators therefore becomes clear when the translator is faced with a situation where he/she can no longer rely on existing knowledge and/or dictionaries, and has to conduct research beyond the dictionary.

There is a clear difference between translating into an international language such as English and translating into so-called 'minor languages' or 'languages of limited diffusion' (LLDs) such as the African languages. This difference also holds regarding the translation of terminology. Cluver (1989: 254) points out that since the terminographer working on a developing language actually participates in the elaboration / development of the terminology, he/she needs a deeper understanding of the *word-formation processes* than his/her counterpart who works on a so-called 'developed language'.

In this paper, a preliminary study is undertaken, comparing and analysing the various translation strategies utilised by African-language translators in the finding of suitable translation equivalents for English terms foreign to the African languages. To this end, a multilingual corpus of ten parallel texts in all eleven of the official South African languages has been studied. These parallel texts have been culled from the Internet, and a full report on the building of this multilingual corpus can be found in De Schryver (2002). The combined size for all eleven parallel corpora is 348,467 running words, or thus nearly 32,000 words on average per language.

2. Methodology followed

The first step in this pilot study is to extract the relevant terminology and to compare the English terms with their translation equivalents in the nine official African languages, *viz*. isiNdebele, siSwati, isiXhosa, isiZulu, Xitsonga, Setswana, Tshivenda, Sepedi and Sesotho, as well as with Afrikaans. For the purposes of this study, we assume that the English texts are the source texts, as all of the websites from which the parallel texts were downloaded, have been written in English, with only small selected sections of the sites in question being provided in the other official languages. Furthermore, it is standard practice in South Africa when undertaking a translation

project involving all nine of the official African languages, to provide the source text in English, as this is in the majority of cases the only language that all of the translators have in common. This is especially the case when the subject matter of the text in question is of a technical nature, as the African languages do not as a rule possess the requisite terminology. On the basis of this evidence it is therefore highly unlikely that any of the African languages (or for that matter Afrikaans) would have served as the source for the texts culled from the Internet on which this study is based.

In extracting the terminology from this corpus of parallel texts, the methodology illustrated by Taljard & De Schryver (2002) is followed. These researchers have shown how African-language terminology can successfully be extracted semi-automatically from untagged and unmarked running text (texts culled from the Internet are, when saved as text files, an example of this) by means of basic corpus query software like *WordSmith Tools*. The key procedure for identifying terminology in each of the parallel corpora is to compare the frequency of every distinct word-type in each parallel corpus, with the frequency of the same word-type in respective reference corpora – the reference corpora obviously being the bigger of the two in each case. Items displaying a great (positive) disparity in frequency are identified as terminology, since the disparity would imply that those specific items occur with unusual (high) frequency in the smaller corpus. The terminology retrieved in this way across the parallel corpora compares very well (see also Uzar & Walinski 2000). For the purposes of this study, the eleven general corpora compiled in the Department of African Languages at the University of Pretoria have been used as reference corpora (for more details, cf. Prinsloo & De Schryver 2002: 256). Sizes of these corpora are typically from a few million up to more than 10 million running words each.

The next step is then to identify the various translation strategies utilised by the different translators in finding suitable translation equivalents for the English terms identified.

3. Preliminary results

On studying the various outputs from the keyword searches done using the two sets of eleven corpora, i.e. the eleven parallel corpora versus the eleven general corpora, the following is readily apparent:

- Although the number of keywords thrown up semi-automatically differs from language to language, there is a good correlation across the parallel corpora between the terms obtained in this manner.
- Even at a casual glance, the following strategies utilised in the translation of source text (ST) terminology are immediately obvious:
  - Translation by means of loanwords in which the English spelling has been retained. Such words have not been transliterated, i.e. nativised in the sense that their phonology has been adapted to reflect the phonological system of the borrowing language.

- Term formation through transliteration. New scientific and technical terms are formed via a process of transliteration by adapting the phonological structure of the loanword to the sound system of the borrowing language.

The occurrence of these two translation strategies in the various languages is summarised in Table 1.

Table 1: Keywords and the translation strategies pertaining to loanwords in eleven parallel corpora

Language	**Keywords**	**Loanwords with English spelling**		**Transliterations**	
	#	#	%	#	%
isiNdebele	583	14	2	37	6
siSwati	427	18	4	16	4
isiXhosa	580	27	5	32	6
isiZulu	619	71	11	30	5
English ST	443				
Afrikaans	426	10	2	17	4
Xitsonga	402	37	9	56	14
Setswana	436	26	6	32	7
Tshivenda	394	43	11	55	14
Sepedi	371	18	5	32	9
Sesotho	320	10	3	13	4

The most important findings regarding these two translation strategies are:

- Whereas isiZulu seems to make use of non-nativised loanwords to a larger extent than transliterations, and whereas siSwati uses these two strategies in equal measure; in all the other languages, i.e. isiNdebele, isiXhosa, the Sotho languages (Setswana, Sepedi and Sesotho), Tshivenda, Xitsonga and Afrikaans, transliterations seem to be used to a greater extent than non-nativised English loanwords as preferred translation strategy for technical terms.
- Many of the non-nativised loanwords under discussion here, are in fact English abbreviations such as SAQA (South African Qualifications Authority), NSB (National Standards Body), RPL (Recognition of Prior Learning), etc. that have been taken over as such into the borrowing language. In Sepedi and Sesotho for example, 78% and 70% respectively of the non-nativised loanwords are English abbreviations that have not been translated into the language concerned, but taken over as is.
- In Afrikaans, translation equivalents are given for English abbreviations such as Eng. SAQA : Afr. SAKO (*Suid-Afrikaanse Kwalifikasie Owerheid*), Eng. NSB : Afr. NSL (*Nasionale Standaardeliggaam*), etc., with the noted exception of the abbreviation ANC (African National Congress).

- A similar situation is found in isiNdebele, where translation equivalents are provided for abbreviations such as: Eng. NSB : Ndeb. iHTB (*iHlangano yesiTjhaba yamaBanga*); Eng. SAQA : Ndeb. iPSAF (*UbuPhathimandla beSewula Afrika*), etc.

4. An illustrative example: comparing translation strategies utilised in isiZulu and Sepedi

As was stated at the outset, this paper is intended as a preliminary investigation into strategies utilised in the translation of terminology into *all* South African languages. As this is an ambitious and wide-ranging project, and as time is limited in a forum such as this, two languages, viz. isiZulu and Sepedi, are used as an illustrative example of this process. In Table 2, a representative sample of 20 SL terms are selected from our large database currently under construction, and this is followed by a comparative analysis of the strategies used in the translation of terminology into these languages.

Table 2: Comparative analysis of 20 SL items translated into isiZulu and Sepedi

SL term	isiZulu translation equivalent	Sepedi translation equivalent
accreditation	**PAU**: *ukunikezwa amandla / igunya*; *ukugunyaza* **BT**: to be given the power / authority, security; to authorise.	**MGW**: *netefatšo* **BT**: verification; **MGW**: *tumelelo* **BT**: permission, approval.
agenda	**PAU**: *uhlelo / uhlu lokuzoxoxwa ngakho* **BT**: arrangement, list of things (issues) that will be talked about / discussed.	**MGW**: *lenaneo* **BT**: list, programme.
apartheid	**MGW** & **MNW**: *ubandlululo (ngokwebala)* **BT**: discrimination (on the basis of colour), exclusion.	**LWT**: *aparteiti* **BT**: apartheid; **MGW**: *kgethollo* **BT**: separation, segregation - **SYN**.
assessment criteria	**PAR**: *inqubo yokuvivinyisisa / yokuvivinya* **BT**: criteria (lit. procedure, process) of examining / examining thoroughly; **PAR**: *indlela yokuhlola* **BT**: manner of examining. (All of these paraphrases are rather vague and do not succeed in capturing the exact meaning of the SL term.)	**PAR**: *mokgwa wa tekanyetšo* **BT**: way / manner of estimation; **PAR**: *dinyakwa tša tlhahlobo* **BT**: requirements of examination.
census	**LWE**: *i-census*; **MGW**: *ubalo* **BT**: count (n) - **SYN**.	**MGW**: *palo* **BT**: count (n.).
definitions	**MGW**: *izincazelo* **BT**: explanations.	**MGW**: *dithlalošo* **BT**: explanations, meanings.
documentation	**MGW**: *izincwadi* **BT**: letters, books; **MGW**: *amabhuku* **BT**: books.	**LWT**: *ditokumente* **BT**: documents.
finance / financial	**MGW**: *izimali / wezimali* **BT**: money / of money.	**MGW**: *(wa / tša) tšhelete* **BT**: (of) money.
gender	**RTE**: *ubulili* **BT**: gender.	**RTE**: *bong* **BT**: gender.
global	**PAU**: *umhlaba wonke jikelele* **BT**: the whole earth, world.	**PAU**: *lefase ka bophara* **BT**: the world at large.

guidelines	**COM**: *imihlahlandlela* **BT**: < *-hlahla* 'guide' + *(i)ndlela* 'way, manner' (Note that the same term is also used to designate 'framework'.); **COM**: *imikhombandlela* **BT**: < *-khomba* 'show' + *(i)ndlela* 'way, manner'.	**COM**: *methalohlahli* **BT**: < *methala* 'lines' + *hlahla* 'guide'; **COM**: *ditšhupatsela* **BT**: < *šupa* 'show' + *tsela* 'road, way'.
institutions	**SSP**: *izikhungo* **BT**: (lit.) gathering places.	**LWT**: *diinstithušene* **BT**: institutions.
Minister	**RTE**: *ungqongqoshe* **BT**: minister.	**CST**: *tona* **BT**: advisor to the chief / king.
outcome(s)	**MGW**: *imiphumela* **BT**: results; **RTE**: *impumelelo* **BT**: outcome, success.	**MGW**: *dipoelo* **BT**: results.
redress	**MGW**: *ukulungisa* **BT**: to correct, rectify.	**MGW**: *phetolo* **BT**: change, reversal.
regulation(s)	**COM**: *imithethonkambiso* **BT**: < *imithetho* 'laws, rules' + *(i)nkambiso* 'custom'.	**MGW**: *melawana* **BT**: small laws.
research	**RTE**: *ucwaningo* **BT**: research.	**LWT**: *resetšhe* **BT**: research; **SSP**: *nyakišišo* **BT**: investigation - **SYN**.
South African Qualifications Authority	**LWE**: *i-South African Qualification(s) Authority*; **PAR**: *Isigungu seziPhathimandla sokwengamela iziqu eNingizimu Afrika* **BT**: authorising committee that presides over South Africa's qualifications - **SYN**.	**PAR**: *Bolaodi bja Mangwalo a Thuto bja Afrika Borwa* **BT**: authority of letters of learning of South Africa.
stakeholder(s)	**MNW**: *abathintekayo* **BT**: those affected.	**COM**: *bakgathatema* **BT**: those who take part.
Standards Generating Body	**LWE**: *iStandards Generating Body*; **PAR**: *uMgwamanda eKhiqiza / eYenza amaZinga* **BT**: assembly, congregation, community that (abundantly) produces / makes standards; **RTE** & **LWE**: *uMgwamanda iStandards Generating Body* - **SYN**.	**PAR**: *Lekgotla la Tlhamo ya Maemo* **BT**: council of establishment of standards.

Note that in Table 2, the SL terms are listed as proffered by the keyword search, i.e. in derived or inflected form. However, should a terminology list be compiled, these terms will be lemmatised under their canonical forms. Note also that the following codes are used to symbolise the strategies that are, according to Baker (1992: 26-42), often used by professional translators in solving various types of problems of non-equivalence at word-level:

- **MGW:** Translation by a more general word (superordinate).
- **MNW:** Translation by a more neutral or less expressive word.
- **CST:** Translation by cultural substitution.
- Translation using a loan word or loan word plus explanation (sometimes in brackets):
  - **LWE:** Translation by means of loanwords in which the English spelling has been retained. Such words have not been transliterated, i.e. nativised in the sense that their phonology has been adapted to reflect the phonological system of the borrowing language.

    - **LWT:** Term formation through transliteration. New scientific and technical terms are formed via a process of transliteration by adapting the phonological structure of the loanword to the sound system of the borrowing language.
- **PAR:** Translation by paraphrase using a related word, i.e. paraphrasing by using a direct / ready equivalent of the SL item in the paraphrase.
- **PAU:** Translation by paraphrase using unrelated words, i.e. paraphrasing by not using a direct / ready equivalent of the SL item in the paraphrase.

In addition to the translation strategies listed above, it is well known that translators working into the African languages are more often than not required to create new terms, and should therefore be completely *au fait* with term creation strategies in their particular language. Regarding term formation in the African languages, Mtintsilana & Morris (1988: 110-112) distinguish between *term-formation processes* internal to the language, and borrowings from other languages. They identify a number of term formation processes in the African languages, of which the following appear in Table 2:

- Semantic transfer: This is the process of attaching new meaning to existing words by modifying their semantic content.
    - **SSP:** In the creation of new terms, the most common form of semantic transfer is *semantic specialisation*, i.e. a word from the general vocabulary acquires an additional, more technical meaning.
- **COM:** Compounding. The term is coined by combining existing words.
- **SYN:** Synonym richness of the vocabulary. Although this is not a method of creating new terms, Mtintsilana & Morris point out that the relative abundance of synonyms in African-language vocabularies offers both advantages and disadvantages from a terminological point of view. E.g., a term may be coined for a foreign concept while a transliteration of the foreign term is also in use.

- Lastly, in some cases in Table 2 above, there is no problem of non-equivalence (at word level) between the source and target languages, as the TL possesses a ready translation equivalent of the SL term in question. Such cases are designated with the code **RTE** (ready translation equivalent).
- The code **BT** in Table 2 above, stands for Back-translation.

The data from Table 2 is quantified in Table 3. (Note that in cases where there are two translation equivalents for a particular keyword, each of these equivalents is counted as a half).

Table 3: Quantitative analysis of 20 SL items translated into isiZulu and Sepedi

Translation strategy	isiZulu		Sepedi	
	# terms	% terms	# terms	% terms
Paraphrase:				
• PAR	2	10	3	15
• PAU	3	15	1	5
Term formation strategies:				
• LWE	1.5	7.5	—	—
• LWT	—	—	3	15
• COM	2	10	2	10
• SSP	1	5	0.5	2.5
More general and/or neutral word:				
• MGW	5.5	27.5	8.5	42.5
• MNW	1.5	7.5	—	—
RTE	3.5	17.5	1	5
CST	—	—	1	5
Total	20	100	20	100

The following conclusions can be drawn from Table 3:

- In both isiZulu and Sepedi, translation by a more general and/or neutral word seems to be the preferred strategy, i.e. in a little more than a third of all cases (35%) in the isiZulu sample and approaching half of the cases (42.5%) in the Sepedi sample.
- The next most popular translation strategy in isiZulu would seem to be translation by paraphrase at 25% of the sample.
- This contrasts with Sepedi where term formation is utilised in just over a quarter of the cases (27.5%) as the next most popular translation strategy after translation by a more general word.
- Term formation as translation strategy is found in just less than a quarter of cases in the isiZulu sample (22.5%).
- In Sepedi, translation by paraphrase accounts for another fifth of the sample (20%).
- In only 17.5% of the cases does isiZulu make use of a ready / direct translation equivalent, whilst in Sepedi the remaining 10% of the cases consists of one instance of translation through the use of a ready / direct equivalent, and one instance of translation through cultural substitution.
- The same translation strategy is used in both isiZulu and Sepedi in the translation of the following SL terms: *assessment criteria*, *definitions*, *finance / financial*, *gender*, *global*, *guidelines* and *redress*.
- In a few cases, both isiZulu and Sepedi display synonym richness. This is the case with the SL terms *census*, *South African Qualification(s) Authority* and *Standards Generating Body* in isiZulu, and *apartheid* and *research* in Sepedi.

5. Conclusion

In this paper we have shown how electronic machine-readable corpora can be used in determining the strategies used by professional translators in finding translation equivalents for SL terms. This is a wide-ranging project that will require the participation of researchers from all of the South African languages, and which will on completion provide a wealth of data with numerous practical applications. Apart from the obvious benefits of this undertaking for the fields of translation studies and terminology, the results from this project will provide guidelines to especially African-language translators, confronted with the onerous task of finding translation equivalents for SL terms foreign to these languages.

References

Baker, M. 1992. *In other words: a coursebook on translation*. London: Routledge.

Cluver, A.D. de V. 1989. *A manual of terminography*. Pretoria: Human Sciences Research Council.

De Schryver, G.-M. 2002. Web for/as Corpus: A Perspective for the African Languages. *Nordic Journal of African Studies* 11/2: 266-282.

Mtintsilana, P.N. and R. Morris. 1988. Terminography in African languages in South Africa. *South African Journal of African Languages* 8/4: 109-113.

Prinsloo, D.J. and G.-M. de Schryver. 2002. Towards an 11 x 11 Array for the Degree of Conjunctivism / Disjunctivism of the South African Languages. *Nordic Journal of African Studies* 11/2: 249-265.

Taljard, E. and G.-M. de Schryver. 2002. Semi-Automatic Term Extraction for the African Languages, with special reference to Northern Sotho. *Lexikos* 12: 44-74.

Uzar, R. and J. Walinski. 2000. A comparability toolkit: Some practical issues for terminology extraction. In B. Lewandowska-Tomaszczyk and P.J. Melia (eds.). *PALC'99: Practical Applications in Language Corpora. Papers from the International Conference at the University of Lodz, 15-18 April 1999*: 445–457. (Lodz Studies in Language 1.) Frankfurt am Main: Peter Lang.

Case Studies on Term Entry and Glossary Distribution

Ewald GEHRMANN
STAR Technology & Solutions, a division of STAR Group, Boeblingen, Germany

1. Introduction

Increasing worldwide specialisation results in an increasing demand for specific information. The speed at which users can access information that is up-to-date and validated is considered to be an important factor. But there are still not enough people who have an in-depth knowledge of what is needed inside a professional organisation. There is a lack of training and experience in how to perform a specific needs analysis, little knowledge about how to structure a terminology database, or what the basics of terminology work are, and there are no tools and processes. But the most important fact is the uncertainty about the value of terminology work for a professional organisation.

In the past few years (since 1996) the most important topic in translation has been the usage of Translation Memory tools, the rise in their functionality and their ability to work with as many file types as possible. Terminology remained a secondary issue and industry was extremely hesitant to pay translators for terminology work. The year 2002 saw a revival in the importance of terminology work, based on the abovementioned factors. There is now a general ability to enter and distribute terms over the Web or intranet, thus providing a channel for fast distribution. As the use of hardcopy terminology during the translation process is diminishing, the importance of online terminology is rising exponentially.

2. What functionality is now expected from a professional terminology system, delivering all the needs of different players?

Working in teams spread worldwide means to provide a system not only for terminology distribution but also for terminology management. This includes the possibility of working in virtual teams.

Sales staff, however, require different information from translators. Thus, it is necessary that the system gives options of configuration to adapt the work environment to meet the varying requirements and work processes of the terminologists. Superfluous information is concealed from the user to promote efficiency and increase concentration.

Users, groups or internal cost centres within an organisation should receive specific access and editing rights to selected dictionaries. This provides them with selective access to corporate terminology. In this way, every employee can minimise the time it takes to achieve his/her objective. It cuts their workload significantly – and this translates into lower costs.

3. Terminology management versus terminology communication

Development is located in India; Marketing is in the U.S.; the technical authors are in Germany; and the translators are everywhere. Each division produces new terms every day. They have to be communicated throughout the organisation as quickly as possible so that every employee can "speak the same language". Communication in virtual teams (cf. Figure 1) often is a problem which causes friction / losses. So, when creating terminology in worldwide virtual teams, the key factor is to provide efficient communication between all the persons involved. Modern Terminology Systems should therefore support queries and comments up to every entry to the remote terminologists within the enterprise. Thus terminologists can receive direct feedback regarding the quality and usage of terms and respond immediately.

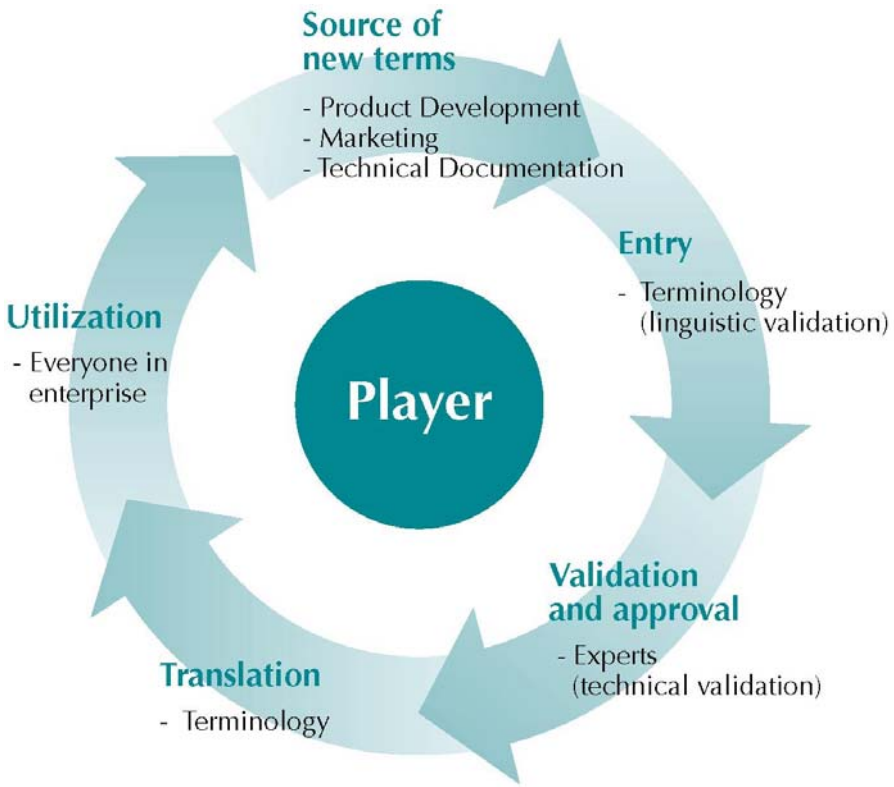

Figure 1: Terminology lifecycle with players of a virtual team

4. Implementation and access?

These days any system's architecture allows rapid terminology updating based on the latest technology. So the new generation of terminology solutions should be easily integrated into existing IT networks, Translation Memory systems, Content Management systems, Office suites or DTP programs, to achieve a high level of consistency and quality when foreign language documents are produced in Documentation, Marketing or Sales.

5. What could be the arguments of the decision-maker?

So what are the facts which will change uncertainty about the value of terminology work for a professional organisation?

- For the General Management: Reduce costs, standardise the global image of the company.
- For Documentation or Translation: Use correct terms in every language, save time-consuming research.
- In Purchasing or Sales: Communicate with customers, strategic partners and suppliers without any misunderstandings.
- In Accounting: No more browsing through dictionaries to check documents.
- In Human Resources: Understanding an application that contains exotic terms and expressions.

During the session, various case studies will be presented. Based on real examples from the software and automotive industry as well as from government bodies, the speaker will describe the background history of various terminology databases. There will be a discussion on how and why global companies have understood that constant terminology work is one of the most important prerequisites for efficient communication and for producing understandable and legally perfect documentation. Terminology is gradually becoming part of a positive corporate identity.

Terminological Shift in a Slippery Economy

Johan D.U. GELDENHUYS
Head of Documentation, Nedcor

Abstract: The terminological shifts between British and American English in the financial sector of South Africa are examined by adducing examples of terms such as *shares and stocks* transmuting to *stocks and bonds* and the impact of this on Afrikaans. In doing so fascinating ramifications are followed to their logical and sometimes not so logical conclusions. Plethoras of terms such as *stocks*, *bonds*, *equities*, *futures*, *forwards*, *options*, *swoptions* and *aandele*, *effekte*, *ekwiteite*, *termynkontrakte*, *vooruitkontrakte*, *opsiekontrakte*, or *ruilopsiekontrakte* are encountered in the process and classified, customised, standardised and harmonised.

During the course of the discussion distinctions are drawn between the funding of companies with their own and with loan capital and the resultant terminological consequences are explored in some detail. The terminological detritus in the evolution of the names of various South African Financial Exchanges, such as *The Johannesburg Stock Exchange*, the *South African Futures Exchange*, the *Bond Exchange of South Africa*, the *JSE Securities Exchange of South Africa* and their evolving Afrikaans equivalents such as *Die Johannesburgse Effektebeurs*, *Die Johannesburgse Aandelebeurs*, die *Suid-Afrikaanse Termynbeurs*, die *Effektebeurs van Suid-Afrika*, and die *JSE Sekuriteitebeurs*, is analysed and explained.

Finally, concomitant terminological fallout in fields such as housing bonds, government bonds and equity accounting is also brought to book and fit into overarching terminological structures to smooth the way for multilingual terminological development in the financial sector of South Africa to the benefit of all its peoples and trading partners.

1. Introduction

There are two legal ways of funding a company: by using its own capital or by using loan capital. In the event of using its own capital a company raises this capital by issuing shares in the common stock of the company. The persons taking up these shares could then receive dividends out of the profits of the company. However, the capital put up by those shareholders remains the company's own capital to be used by it in generating profits. It is not paid back to the shareholders.

In using loan capital a company pays interest or a return on the capital borrowed and has to repay the capital itself within a certain period. If the loan amount is huge, the company can issue stock to the value of for example hundreds of millions of Rand. The stockholders will then receive a fixed yield or return over time (dividends paid to shareholders fluctuate according to profit moves) and the full capital amount at the end of the period or term of the stock. In this instance the company has therefore borrowed the capital and has to repay it.

2. British versus American English terminology

This was the case terminologically speaking in South Africa until very recently. Because linguistically South Africa was regarded as a dominion of Great Britain, we followed the British way in financial terminology and not the American one. However, this has changed dramatically in that even the remnants of Great Britain are now using American terminology when it comes to financial documentation.

What the British referred to as *shares* (short for *shares in the common stock* of a company), the Americans referred to as *stock* (also short for *shares in the common stock* of a company). Therefore they could not refer to *loan stock* as *stock*, but used *bonds* to indicate instruments used in raising loan capital. All very confusing up to a point: *shares and stocks* in British English became *stocks and bonds* in American usage.

3. South African English terminology

In South Africa this change was reflected in the Afrikaans terminology for *The Johannesburg Stock Exchange*, as it used to be known. Initially *The Johannesburg Stock Exchange* was known as *Die Johannesburgse Effektebeurs*. As *aandeel* equates to *share* and *effek* to *stock*, this meant that the British English pattern was reflected in Afrikaans. In the early nineties the *Bond Exchange of South Africa* started business, obviously following the American paradigm with *stocks* equalling British *shares* and hence Afrikaans *aandele*, and *bonds* equating to *stocks* and *effekte*, respectively. So, in Afrikaans the *Bond Exchange* became the *Effektebeurs* and the *Stock Exchange* the *Aandelebeurs*. Therefore the Afrikaans equivalent of *The Johannesburg Stock Exchange* was *Die Johannesburgse Aandelebeurs*. This was not to be the last name change of that august institution, however.

Also in the early nineties, the *South African Futures Exchange* took off, trading mostly in *futures*, but in some *options* too. As *The Johannesburg Stock Exchange* was also trading in *share options* or rather *stock options* by now, there was a little spin-off on the margins, as the Americans would have said. Nevertheless *futures*, *options*, *stocks* and *bonds* now all had definite homes for trading, with *options* just a mite promiscuous financially in that they kept two establishments. This little problem was, however, solved when *The Johannesburg Stock Exchange* and the *South African Futures Exchange* merged in the early two thousands, or is it two hundreds by the British count, to form the *JSE Securities Exchange South Africa*.

4. Afrikaans terminology

Now this was splendid for the English and Americans, not respectively, but Afrikaans terminology once more faced a conundrum in that the Afrikaans for *securities* had been *effekte* all along. Unfortunately that term was taken by the Effektebeurs or Bond Exchange, which did not form part of the merger. Thus the *JSE Securities Exchange South Africa* became the *JSE Sekuriteitebeurs Suid-Afrika* and *securities sekuriteite*, as

it were. The thinking behind this was that *securities* or *sekuriteite* represented the widest term encompassing *stocks*, *bonds*, *equities*, *futures*, *forwards*, *options*, *swoptions*, etc., or *aandele*, *effekte*, *ekwiteite*, *termynkontrakte*, *vooruitkontrakte*, *opsiekontrakte*, *ruilopsiekontrakte*, etc.

Incidentally, since *forwards* had been referred to as *termynkontrakte* in Afrikaans before the advent of the South African Futures Exchange, because *vooruitkontrakte* was frowned on as being too directly reflective of the English, a lot of covert pussyfooting was necessary to accommodate *futures* as *termynkontrakte* and *forwards* as the suddenly acceptable *vooruitkontrakte*. Again the distinction was necessary since, even if they share certain characteristics and in fact are somewhat close as financial instruments or securities, there are financially vital differences between futures and forwards such as that futures are freely tradable on an exchange and forwards not. But that is by the by.

To get back to the JSE Securities Exchange South Africa: whatever happened to the Johannesburg in the name? Is JSE tradable and Johannesburg not? Well, the answer to that is relatively easy in that the exchange is now situated in Sandton and not Johannesburg. Perhaps *The Economist's* recent styling of Sandton as a suburb of Johannesburg is best left alone in a discussion of terminological shift and rather relegated to one on landslip or subsidence in general.

5. Tradable versus non-tradable securities

Futures are easily tradable: forwards not. By the same token bonds are tradable: ordinary loans such as personal loans, leases, instalment credits and mortgages or (housing) bonds (you see where I am going with this one) not. Why not complicate things even further. A *housing bond* or *residential bond* or just plain *bond*, as it used to be known, is a *verband* in Afrikaans and short for *housing* or *residential mortgage bond*. Therefore it would make admirable sense henceforth to style it as *mortgage / verband*. Then there are also the old style government bonds such as *defence bonds* and *bonus bonds*, which in Afrikaans are *verdedigingsobligasies* and *bonusobligasies*. Also these terms will have to be cleaned up if ever again our government is going to issue such bonds. Hopefully, peace and the Lotto will make that unnecessary.

Obviously one of the great advantages of tradable securities is that the putter-up or put-upper (let's not go there) of capital is not locked into his/her or its initial investment in that he/she or it can relatively easily dispose of such securities at any time to raise cash or invest in other securities for so long as they are freely tradable. Thus was invented the process of *securitisation* or *sekuritering* whereby a bunch of untradable loans such as individual housing bonds are securitised into one enormous security, which is then floated on an exchange and freely traded – further proof of the encompassing reach of the term *security*.

6. Equity capital versus ordinary share capital

To get back to the beginning: the term *equity* was used to differentiate between *ordinary shares* or *equity* and *preference shares* or *stock* then and *bonds* now. A *preference share* is a share offering a fixed yield and repaying all its capital at the end of a fixed term – in other words a *loan stock*, with the distinguishing characteristic, however, that the term *preference* indicates that, when it comes to receiving a final cut on for example the liquidation of a company, preference shares are preferred to ordinary shares or ranked above them, with the result that preference shareholders are paid out before ordinary shareholders and therefore more likely to receive something than the holders of equity. The Afrikaans for *equity* was *gewone aandeel* or just *aandeel*, with its concomitant confusion.

When a holding company holds more than fifty per cent of the ordinary shares or equities of another company – say, sixty four per cent – it can *equity account* the profits of its subsidiary. If the subsidiary's profits amount to R100m, the holding company can show R64m in its books as its share of the subsidiary's profits. In Afrikaans *equity account* became *ekwiteitsverantwoord*, with no sign of *gewone aandeel* or *aandeel*. A further complication was introduced when *compulsorily convertible preference shares* or *verpligtend omskepbare voorkeuraandele* were regarded as part of a company's equity, but not of its ordinary shares. This was the case because such shares were a hybrid, sharing the characteristics of a bond, fixed yield, for a time before being compulsorily converted into ordinary shares, with a fluctuating dividend stream and no return of capital at the end of a fixed period. To differentiate between *equity capital*, in the sense of *ordinary share capital* plus *compulsorily convertible preference capital* as well as *compulsorily convertible debenture* (another kind of *bond* called a *skuldbrief* in Afrikaans) *capital* plus, in some instances, a portion or all of the share premium account (don't ask – horribly higher accounting), and *ordinary share capital* or *share capital* Afrikaans was of course forced to use *ekwiteitskapitaal*.

7. Tandem terms

This solved some problems but, as always, raised others, especially in the case of tandem terms such as *shares and stocks, stocks and bonds*, *equities and securities, equities and bonds. Shares and stocks* was the old British version of *aandele en effekte,* whereas *stocks and bonds* has always been the American version of the same, with *equities and securities* a later British import to these shores. So, in a way they all meant *aandele en effekte* (remember a company's own capital versus capital it borrowed or loan capital). From an investment point of view *shares and stocks* made perfect sense from the nineteen sixties to the nineteen eighties, with *equities and securities* putting in sporadic appearances from the nineteen seventies onwards. In the nineties *stocks and bonds* started to take over, but not quite. *Equities and securities* is currently being avoided, because *securities* has become such an all-encompassing term, while *shares*

and stocks is slowly being sidelined owing to developments at the JSE Securities Exchange South Africa and the Bond Exchange of South Africa. From an international perspective *stocks and bonds* would be the simplest, neatest solution, with *aandele en effekte* its Afrikaans equivalent. However, South Africa being South Africa and crisscrossed by many languages, of which British and American English are but two, currently we seem to be favouring the hybrid *equities and bonds* for *aandele* (or *ekwiteite* – the mind boggles) en *effekte*.

8. Outlook

With South African English and Afrikaans now having classified, customised, standardised and harmonised their financial terminology, a new challenge is to develop financial terms in the other official languages of South Africa.

Introducing TshwaneLex – A New Computer Program for the Compilation of Dictionaries

David JOFFE[°], Gilles-Maurice DE SCHRYVER[‡] & D.J. PRINSLOO[#]
DJ Software, Pretoria, SA[°], Department of African Languages and Cultures, Ghent University, Belgium[‡] & Department of African Languages, University of Pretoria, SA[‡ #]

1. Introduction

One would expect that it would be possible for a prospective dictionary compiler to walk into a local software mega-store and to merely buy a dictionary compilation program from the shelf, as is quite a reasonable expectation for finding a word processing package, OCR software or even voice recognition tools. This is however not the case. In most set-ups dictionary compilers and publishing houses all use either their own-compiled, locally networked, closed systems, or dictionary compilation is simply done on a word processor such as Microsoft Word or Corel WordPerfect.

No sophisticated SA-designed dictionary compilation software is available either on the local market or internationally. This is a major headache not only for individual dictionary compilers in South Africa but also for this country's *National Lexicography Units* (NLUs) tasked with the compilation of monolingual and bilingual dictionaries for the nine official African languages of South Africa. At the moment these units are left with no choice other than to compile their dictionaries with word processing software until such dedicated software becomes available, preferably a package adaptable to dictionary compilation for any language.

In this interim period some dictionary compilers are smart enough to *simulate a database* while working in a word processor by designing artificial records and fields with non-printing labels such as *lemma*, *definition*, *examples of use*, *part of speech*, etc. This is simply done by creating a template using hard page ends as the beginning and end points of a dictionary entry, simulating a record, and hard paragraph breaks for field delimiters. Thus all dictionary entries compiled this way have the same structure. Although this is a far cry from the real thing, it increases the potential success rate of eventual importation into a new dictionary program substantially.

In the past few years many individuals and institutions worked hard under the auspices of the *Pan South African Language Board* (PanSALB) to get the units established and up and running and to guide them on a continuous basis. However, a computer program for the compilation of dictionaries is still the crucial missing link in the entire lexicographic process. Initiative was taken in 1998 by D.J. Prinsloo to facilitate the compilation of such a program by bringing together computer experts, lexicographers and publishers in a so-called *SA Invitation Team* – an effort later abandoned in favour of the Swedish program, *Onoma*. When Onoma was discontinued

one year ago, a team consisting of the current authors quickly set out to create a brand-new and innovative software package for dictionary compilation.

It was envisaged that the development of a new sophisticated program for dictionary compilation should take at least five years, but the first version, *TshwaneLex 1.0*, is already imminent. TshwaneLex has its roots in Africa, appropriately linking *Tshwane* 'Pretoria' and its main function, *lexicography*, in a single acronym. TshwaneLex will also be marketed abroad. Dictionary compilers in the UK, Europe and Australia already show a keen interest in this program. Commercial aspects such as licensing of the software and conditions as well as methods of purchase still need to be finalised.

TshwaneLex is a software tool for all-purpose dictionary compilation, with some specific customisations for the African languages. The initial primary focus of development is to fulfil the requirements of the NLUs. The first version of TshwaneLex will not have built-in corpus tools or terminology management functionality. These capabilities will be considered for future versions of the software.

Some of the general 'guiding design principles' for TshwaneLex are:

- *User-friendly*: The compilation environment is intended to be as intuitive as possible, in order to minimise required training time and required computer literacy level.
- *Speed*: TshwaneLex is intended to provide the most 'streamlined' possible editing experience for lexicographers. All primary editing tools should be quickly accessible from the main interface, as well as by means of keyboard shortcuts.
- *Immediate output preview*: There should be an immediate preview showing how the article would appear in a printed dictionary.

Initial practical testing and development of the software is being done on a bilingual dictionary as well as a monolingual dictionary. SeDiPro, a bilingual *Northern Sotho – English* dictionary of the Department of African Languages at the University of Pretoria that was originally compiled on a word processor, was successfully 'transferred' to TshwaneLex and is currently being expanded with TshwaneLex. Monolingual Northern Sotho (Sesotho sa Leboa) data is used to test the use of TshwaneLex in the compilation of a major monolingual dictionary.

2. Overview

TshwaneLex can be used for the compilation of both monolingual and bilingual dictionaries. The user is presented with a 'language editing window' for each language in the dictionary. This language editing window provides the primary interface for editing the dictionary.

The language editing window consists of four primary parts, as shown in Figure 1.

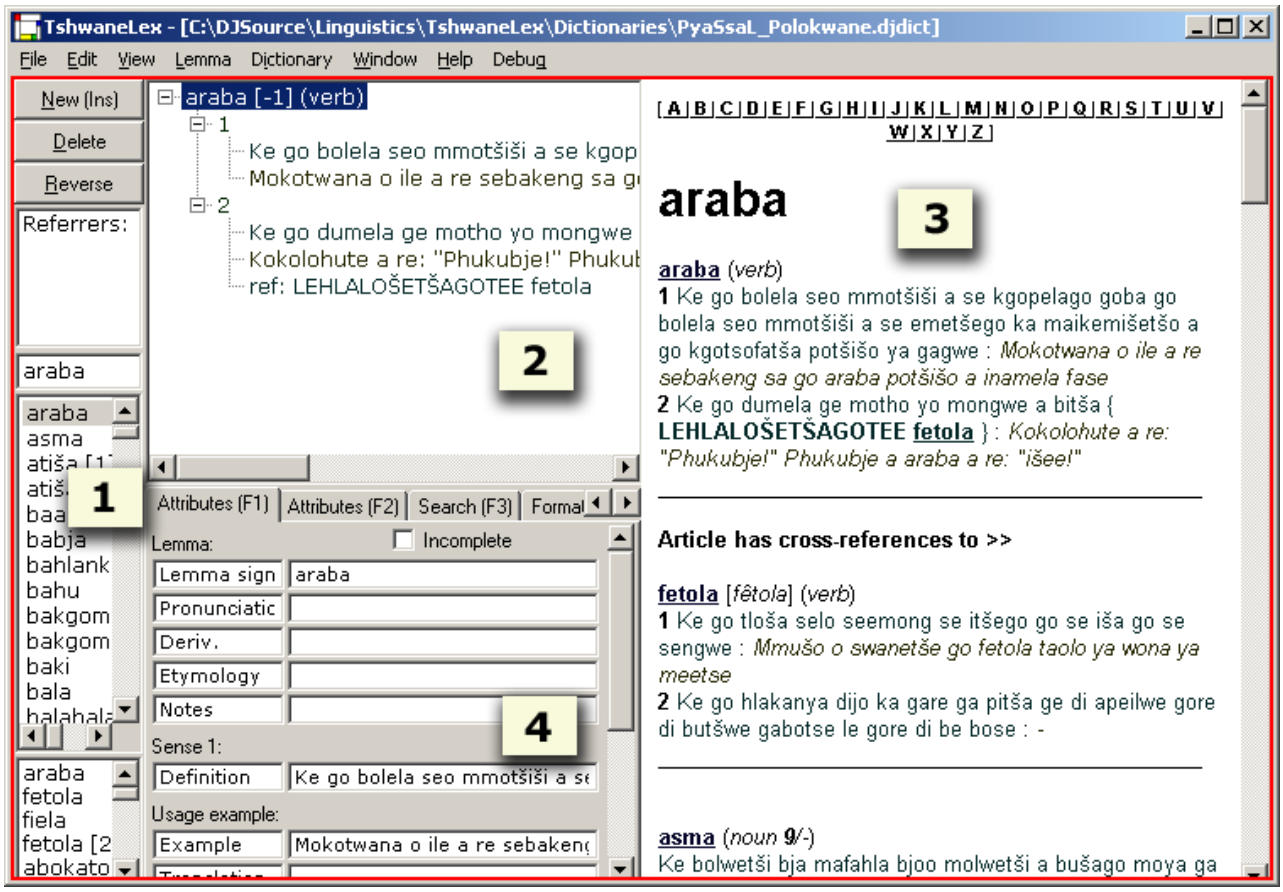

Figure 1: TshwaneLex screenshot showing the four main editing windows

1. **The lemma list**. A list of all lemmas for the language is shown at all times, from which the lexicographer can select which lemma to view or work on.
2. **The tree view control**. The 'tree view' shows the hierarchical structure of the currently selected article. This indicates the hierarchical relations of all senses, sub-senses, definitions, translation equivalents, usage examples, cross-references and combinations in the selected lemma. The tree view control also allows the structure to be edited, for example word senses may be created or deleted or moved around. Each element belonging to the selected article is represented by a 'node' in the tree. The specific text fields associated with that node are attributes of that node, and are edited in the tools window (cf. main editing window 4.).
3. **The preview area**. The primary purpose of the preview window is to show a preview of the article that closely resembles the output that would appear in a printed dictionary. Different elements of the article are displayed in different colours, for easy visual distinction. The article preview updates immediately whenever any changes are made to the article. One very useful feature of the preview area is that all cross-references related to (i.e. cross-references to and from) the currently selected lemma are also listed. The lemma signs of all cross-references

are displayed as clickable hyperlinks, allowing the referenced articles to be selected quickly.

4. **The tools window**. The tools window consists of a number of sub-windows. These are used for modifying attributes of an article, and provide other tools such as a text search function and a lemma 'filter'. These will be explained in more detail in §3.

When in bilingual editing mode, there are two language editing windows, one for each language. The two windows are arranged side by side, splitting the screen into two parts down the centre. This allows the lexicographer to work on both languages *simultaneously*. When working on a lemma in the source language, the lexicographer can immediately see how the translation equivalents of that lemma are treated in the target language, and vice versa.

3. The tools window

The tools window consists of five functionally distinct sub-windows, which are described in the following five sections.

3.1. Text attributes

This window contains text controls that allow the lexicographer to edit any of the text fields of the currently selected lemma. Depending on where one stands in the tree view, different (and appropriate) text controls are automatically generated.

Incomplete tag

This window also contains a checkbox labelled 'incomplete'. If a lexicographer is unsure about some aspect of the treatment of an article, and wants to return to it later, this checkbox can be used to tag the lemma as being incomplete. Lemmas that are marked as incomplete are automatically excluded from the output when the dictionary is being prepared for printed form.

3.2. Usage labels, parts of speech, noun classes

This window allows the lexicographer to specify the following attributes:

- usage labels (e.g. *formal*, *offensive*, etc.);
- part of speech tags (e.g. *noun*, *verb*, etc.);
- noun classes.

These may be applied to the lemma or to specific word senses within a lemma.

3.3. Search

This function allows the user to search the dictionary for a text string. Standard search options such as 'whole word only' and 'case-sensitive' may be selected.

3.4. Format and preview display options

An important feature of TshwaneLex is that the language of the metalanguage can be set. This simply means that the language for cross-reference types (e.g. *see*, *compare*, etc.), parts of speech and usage labels can be changed throughout the dictionary with just one mouse-click. This allows multiple dictionaries to be created from the same dictionary database for different markets. For example, two different versions of a bilingual Northern Sotho – English dictionary can be produced; one for English speakers, and the other for Northern Sotho speakers.

3.5. Filter

The filter allows the lexicographer to define criteria for selecting and viewing a subset of the articles in the dictionary. For example, the lexicographer can select to view "only those lemmas that are marked as incomplete", or to view "only those lemmas that do not have usage examples".

Fairly complex filters can be defined by combining the "include/exclude" filters and the "and/or" options. For example, in Figure 2, a filter is defined to show all articles that have cross-references but not definitions.

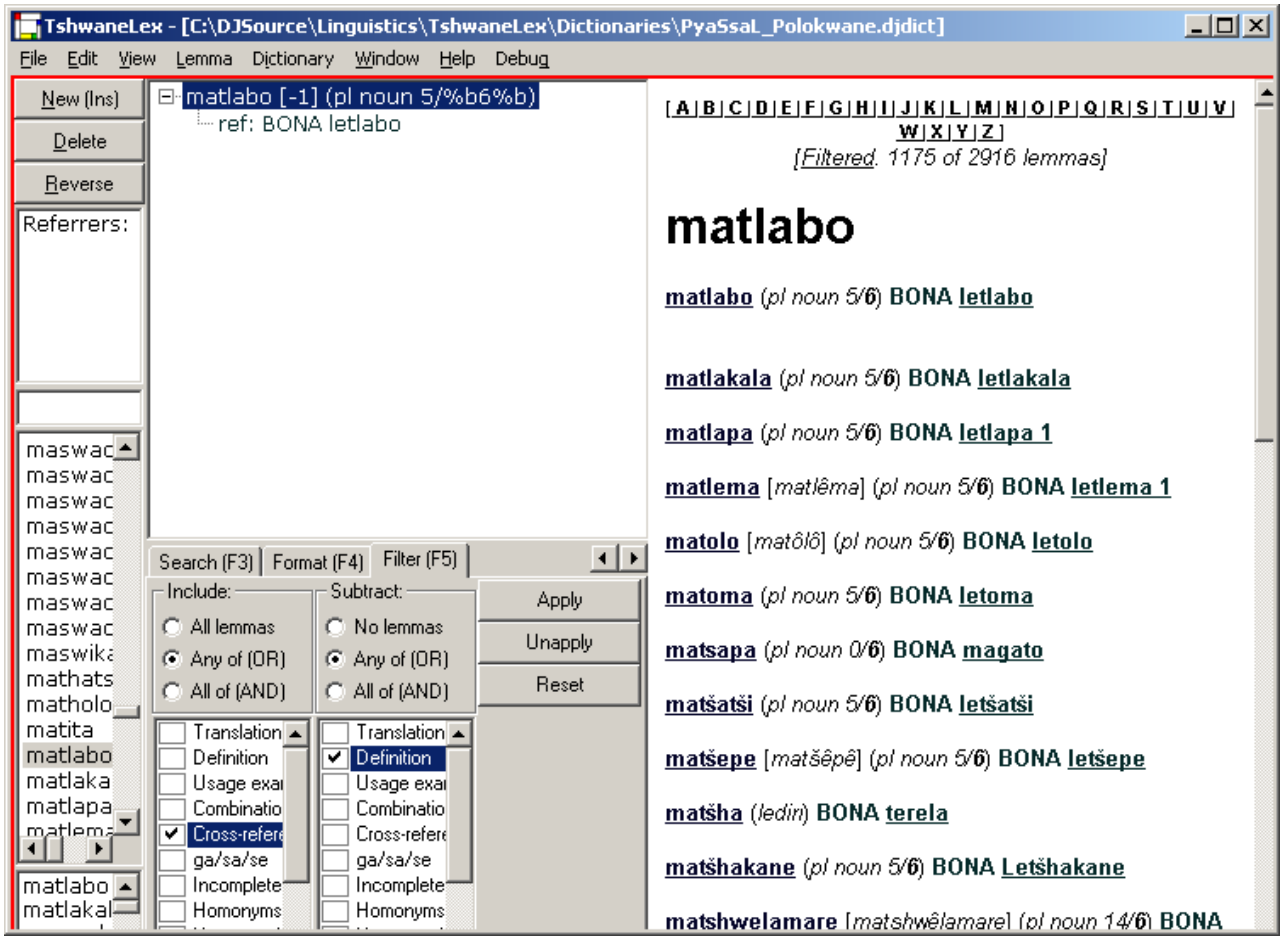

Figure 2: A screenshot demonstrating the 'filter' function (F5)

4. Bilingual editing features

4.1. Linked view mode

When in 'linked view mode', the language window for the target language automatically displays related lemmas when a lemma in the source language is selected, and vice versa. This is highly useful for quickly comparing the treatment of equivalent words on both sides of the dictionary.

4.2. Automatic reversal

TshwaneLex provides functionality to automatically reverse lemmas; that is, to automatically generate reversed lemmas in one language from the other. Either a single lemma can be reversed, or a full reversal can be done on all the lemmas in a language. The full language reversal has various options for how to handle multi-word translation equivalents. The lexicographer may choose to reverse single-word lemmas only, up to two-word lemmas, up to three-word lemmas, or the full dictionary.

Naturally, automatic lemma reversal is not all that is required to generate the treatment of words in the reverse side of the dictionary. Normally the lexicographer will still have to do work on these automatically created lemmas. For this reason, all lemmas generated by automatic reversal are marked as incomplete, indicating that they require further work from the lexicographer. Automatic reversal is thus not intended to *replace* manual reversal, but rather it is intended merely as a tool to speed up the lexicographer's work.

5. The cross-reference system

Automatic sense and homonym update

Internally, TshwaneLex stores the actual *structure* of cross-references, rather than storing a cross-reference as a 'dumb' text string. TshwaneLex uses internally unique identifiers for each referenced lemma or sense 'node'. This allows the system to automatically update the cross-reference should the target lemma sign, homonym number, or sense number of the cross-reference target change.

As an example, if the verb *araba* 'answer' contains a cross-reference to "**fetola :2**" (*fetola*, sense 2), and a lexicographer working on *fetola* decides to switch senses 1 and 2 around, then the cross-reference from *araba* will automatically update itself to "**fetola :1**" (*fetola*, sense 1).

6. Dictionary error checks

TshwaneLex can search for a variety of dictionary errors that lexicographers may make, such as duplicate definitions, duplicate translation equivalents, duplicate combinations, bracket nesting errors, redundant cross-references, redundant white-space and more.

7. Dictionary compare/merge

The dictionary compare/merge function allows a dictionary database to be compared to another dictionary database, with the possibility to merge new or modified lemmas into the main dictionary database. This functionality is particularly useful for situations in which lexicographers are split up geographically, and a high-speed network connection to the primary database server is logistically or economically infeasible. This is usually the case in all but the most developed countries. This feature allows the work done by remote teams to be periodically merged into the primary database.

This feature is also useful for comparing a dictionary against an older, previously backed up, version of itself; the lexicographer can thus easily see what changes have been made since the backup. This function is illustrated in Figure 3.

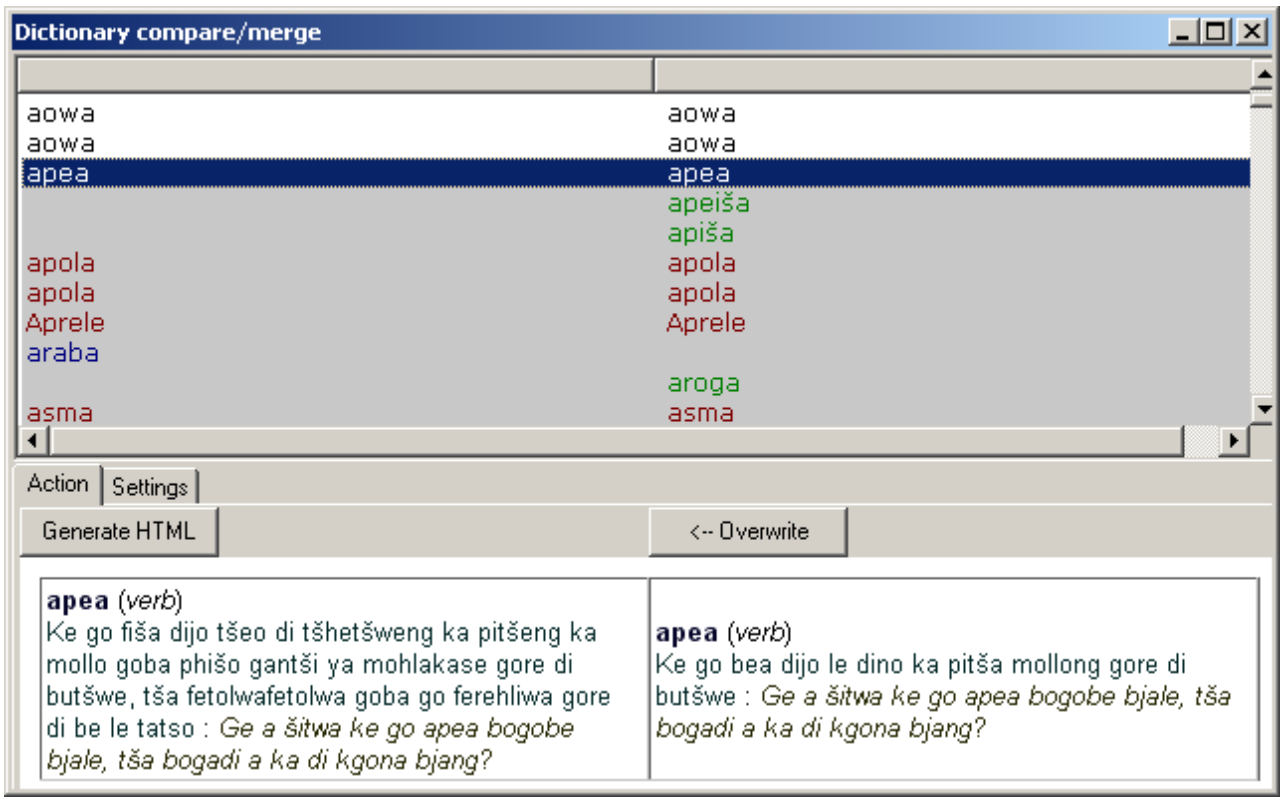

Figure 3: A screenshot of the dictionary compare/merge window

8. Backup system

Regular database backups are an absolutely essential part of any dictionary project. For this reason, TshwaneLex has some built-in functions to assist with the process of creating backups. This includes an option to automatically create backups at specified intervals.

9. Database interface and importers/exporters

TshwaneLex interfaces with a relational database server to access and modify the dictionary database. Multiple lexicographers can thus connect to the server and work on the dictionary simultaneously.

TshwaneLex uses the ODBC (Open Database Connectivity) cross-platform database interface standard to connect to the database server. This essentially allows TshwaneLex to use any ODBC-capable database software. ODBC drivers are available for all major database products on the market.

TshwaneLex can import wordlists and word frequency lists generated by corpora tools.

TshwaneLex has the following built-in exporters:

- XML (eXtensible Markup Language)
- RTF (Rich Text Format), which can be loaded into Microsoft Word
- Static HTML (HyperText Mark-up Language)

Some possible exporters and interfaces planned for future versions include the LaTeX typesetting system, a dynamic web page interface (for online dictionaries), as well as a file format and software interface for electronic dictionaries (e.g. CD-ROM based).

9.1. Custom importers/exporters

TshwaneLex has been built with a modular design that allows additional custom input/output interfaces to be created, should a particular user want to use something other than ODBC. This modularity allows customized importers to be developed should it be required to *migrate* an existing dictionary database from another system. Custom dictionary export (output) modules could also be developed.

References

Further information on acronyms/technologies mentioned in this paper is available at the following locations:

HTML. *HyperText Mark-up Language*. <http://www.w3.org/MarkUp/>

LaTeX. http://www.latex-project.org/

ODBC. *Open Database Connectivity*. <http://www.microsoft.com/data/odbc/default.htm>

RTF. *Rich Text Format*. <http://msdn.microsoft.com/library/default.asp?url=/library/en-us/dnrtfspec/html/rtfspec.asp>

XML. *eXtensible Markup Language*. <http://www.xml.org/>

The Evolution of Version 2 – A Multilingual Database for A Multitude of Users

Barbara I. KARSCH
J.D. Edwards, Denver CO, USA

Abstract: Since December 1999, the terminology group at J.D. Edwards, a leading developer of agile software solutions, has been working on a multilingual database. First, we designed the software; then we started populating the actual database. We also set process and language standards for the source language, English, and the 19 target languages. Furthermore, we created and delivered training modules for approximately 50 translators and a small number of users in other departments. In recent months, demand for this solution has risen within the company. This paper will discuss the requirements phase for Version 2 of this solution. During this phase, the tool is undergoing changes, and some of the current entry structures and standards are being re-evaluated as well. This paper discusses the changes that have already become apparent. The presentation will describe the set of requirements that will have emerged by February 2003.

1. Introduction

By the end of 2000, the design and implementation phase of the J.D. Edwards terminology tool was complete, and translators, terminologists, and subject-matter experts began populating the database in 2001 (Madsen et al. 2002). In November 2002, the database contained terms, definitions, contexts, and other information for almost 4,000 English concepts and approximately 42,000 entries in 19 target languages (Arabic, Czech, Danish, Dutch, Finnish, French, German, Italian, Japanese, Korean, Norwegian, Polish, Portuguese, Russian, Simplified Chinese, Spanish, Swedish, Traditional Chinese, and Turkish). It also contained simple ontologies for some of the most important conceptual areas of the J.D. Edwards products. A year and a half after the launch of Version 1, there is unexpectedly high demand for both the content and the software within the company.

Demand does not arise only from translators of recently added languages, such as Arabic, or translators and terminologists whose language needs were not met sufficiently in Version 1. Demand also comes from departments such as Marketing and Product Management, and J.D. Edwards foreign offices.

2. The content

To meet the information needs, the terminology team has been working diligently to populate the database. J.D. Edwards currently employs seven full-time terminologists for the seven main languages: Brazilian-Portuguese, French, German, Italian, Japanese, Spanish, and Simplified Chinese. During the summer of 2002, five terminology interns temporarily joined our ranks, a first for the terminology group, adding extra hands to the task. Two of these interns, who were native English speakers, helped create and

edit English language entries. Interns for Italian, Japanese, and Brazilian-Portuguese were tasked with transferring language entries from legacy databases.

3. The tool

Demand for content can only be met, however, if the appropriate tool is available. TDB Version 1 is a client/server software application that is written in Microsoft Visual Basic and resides on a SQL Server database. It comprises 50 tables, version control, an integrated workflow with several quality assurance steps, sufficient security for the translation department, and the basic functionality to provide one-dimensional ontologies.

To make content accessible to users as far away as China or Brazil or to provide the capabilities required by the J.D. Edwards Marketing department, a new Web-based version is essential. As of August 2002, the requirements phase for the new version is underway.

4. Foreign offices

The J.D. Edwards offices and business partners outside the United States need a more solid solution. Today, terminologists send database extracts of their language team's glossary abroad – in either Microsoft Access or Excel format – to help trainers and consultants with their daily work. The requirements from these users include a simple look-up tool as well as term-request or revision-request functionality. A beta version of this interface and search engine exists already, and the first version is scheduled to go live in early 2002 on the J.D. Edwards intranet.

5. Software engineering

While we still have not reached the entire Software Engineering Department, there are pockets within the organisation that recognise the value of a terminology database (Wright & Wright 2001). The Product Management team has expressed interest in sharing the database since software engineers and product managers duplicate much of our research efforts. One of the apparent discrepancies is the different approaches the two departments take: Software engineers are solely interested in concepts expressed in functionality and don't give much thought to their linguistic designators; translators look at these expressions first and try to derive the concepts from them; poorly designated concepts force translators to do time-consuming research into the functionality.

A modification of Frege's semantic triangle in Figure 1 shows how terms, concepts and functionality are related (Karsch & Pichler 2000).

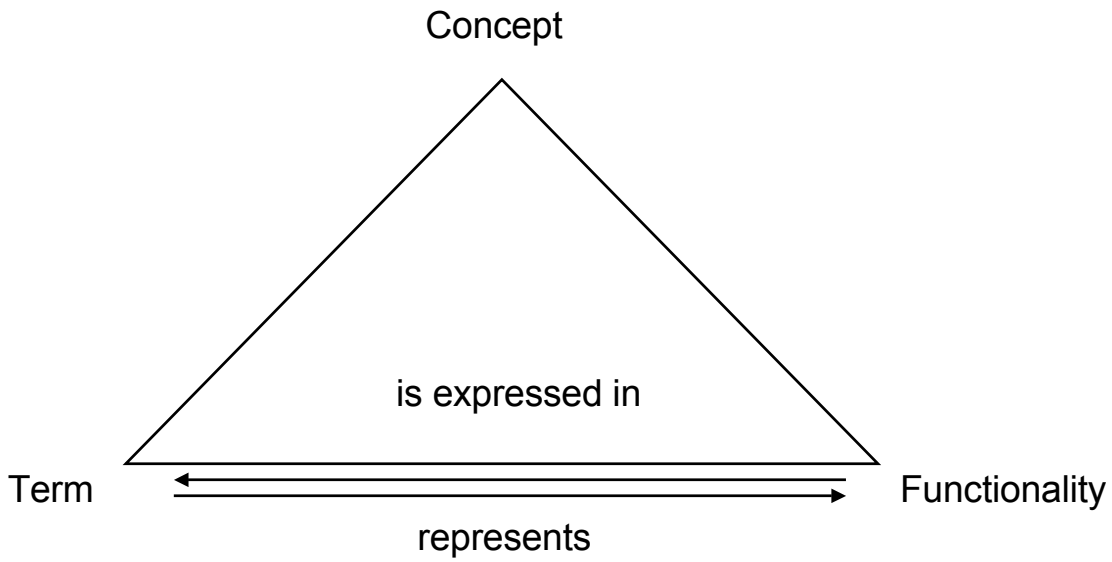

Figure 1: Software functionality and Frege's triangle

It can therefore be expected that the terminology database could and should bridge this gap. If software engineers worked with an ontology and a large conceptual database, they could find good designators to express the functionality they need to name. Figure 2 shows how translators and terminologists can already today query the database for a term, find the concept and add the equivalent in the target language (Karsch & Searle 2001). The advantage here is that they can be prescriptive in language.

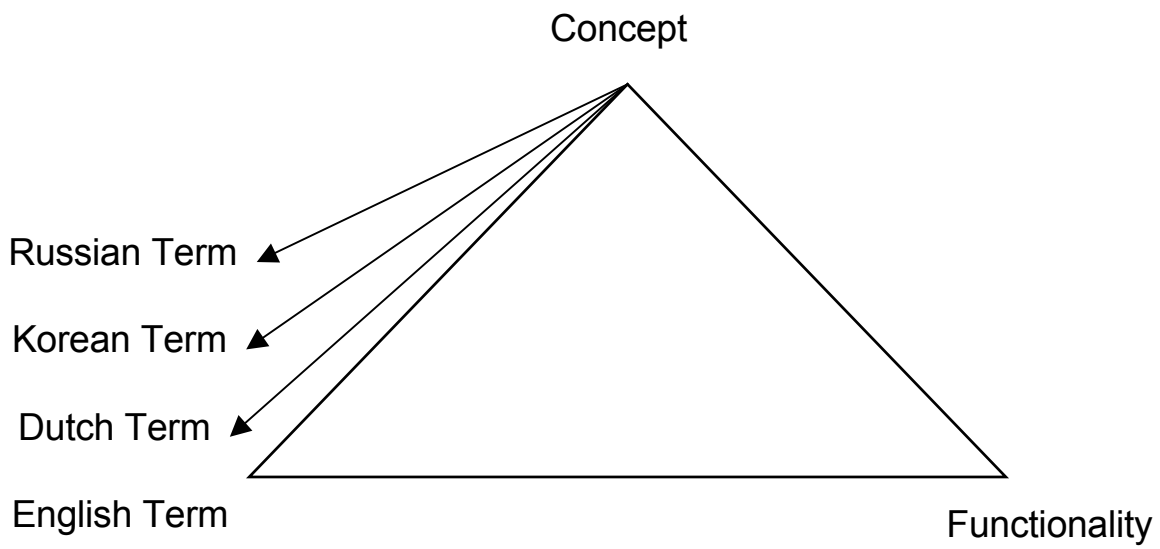

Figure 2: Designators in various languages and Frege's triangle

Currently, translators use the database just as described above. But in fact software developers cannot make full use yet. Descriptive definitions in the terminology database focus mostly on the general meaning of a term rather than explaining what concept underlies the functionality expressed by a term in the software. This was the

primary concern for the Translation department. All software engineers can do today is check for linguistic information and the definitions and make better term choices. The terminology database can only function fully when it is prescriptive. Prescriptive terminology is also the only real foundation for an ontology, which in turn would make automatic term generation possible or at least provide help with the naming of new concepts. For a formal ontology and automatic term generation, more formal devices or description logics (Ceusters et al. *in press*) would be necessary. So far we are technically not ready to tackle such a large project. Furthermore, it would require executive support, which may be easier to attain given our recent collaboration with the Marketing department.

6. Marketing

A Marketing department plays an important role in any organisation. This is where corporate language translates into hard currency. Product names that should be used uniformly throughout the organisation are primarily marketing-driven. The keeper of these product names has recently turned to the terminology group for a database solution. For her, terminology management means not only properly recording product names, but also tracking dates of inception or obsolescence, product owners within the company, marketing descriptions, smallest selling units, and software prerequisites. The Marketing department will also greatly benefit from TDB's ontology functionality: The product structure is highly hierarchical, and every sub-product belongs to a larger group in which it is sold to customers, for example.

Country offices, Software Engineering, and Marketing are new customers, and their needs will be mostly along the lines of look-up capabilities that do not affect the current solution; however, the current user groups have already presented additional requirements and will undoubtedly come up with more.

7. Translations

In some cases, more will actually be less. Version 1 was a trial version where we experimented with many different functions and solutions. Version 2 will, if not do away with certain functions altogether, at least hide them for the users who don't need them. Requirements from the Scandinavian translators, for example, reinforced the demand for a simpler and user-friendlier interface. While the current version does not allow every user to use every possible function, it displays all capabilities to all users. As a consequence, the learning curve is longer, the need for retraining higher, and the potential for errors greater.

Stakeholders of the database rely on a variety of reports. The terminologists, as the guardians of quality, are interested in data integrity. The Translation managers keep track of the quantitative progress, and the translators are interested in their individual contributions. So far, a fairly rich array of reports in Microsoft Access has fulfilled these requirements. However, reporting authority was restricted to one person in order

to avoid involuntary changes to the database. For Version 2, the reporting function will become part of the tool.

Other user groups that were not represented sufficiently during the first requirements phase needed to gain experience in terminology management before deciding upon their requirements. Arabic has just been added to the language repertoire that is handled in-house at J.D. Edwards, so it might still be too early to fully explore the needs of this language. Asian terminologists, representing Japanese, Korean, and traditional and simplified Chinese, on the other hand, have been able to test this terminology environment and have come up with their own subset of requirements.

One proposal that evolved partially in reaction to complaints from the Asian group has been to make terminology management more context-driven (Collet 2002). What we envision is an entry structure in which the English definition and the context of a particular representation of the term (e.g. the full form) are linked to one representation of the concept in a foreign language. That is to say, the English entry that is attached to the concept contains all the abbreviations and acronyms, but remains a single entry. This entry will contain several examples, each representing a different context of the term and using as many representations as possible.

There are several benefits to this structure, which is depicted in Figure 3.

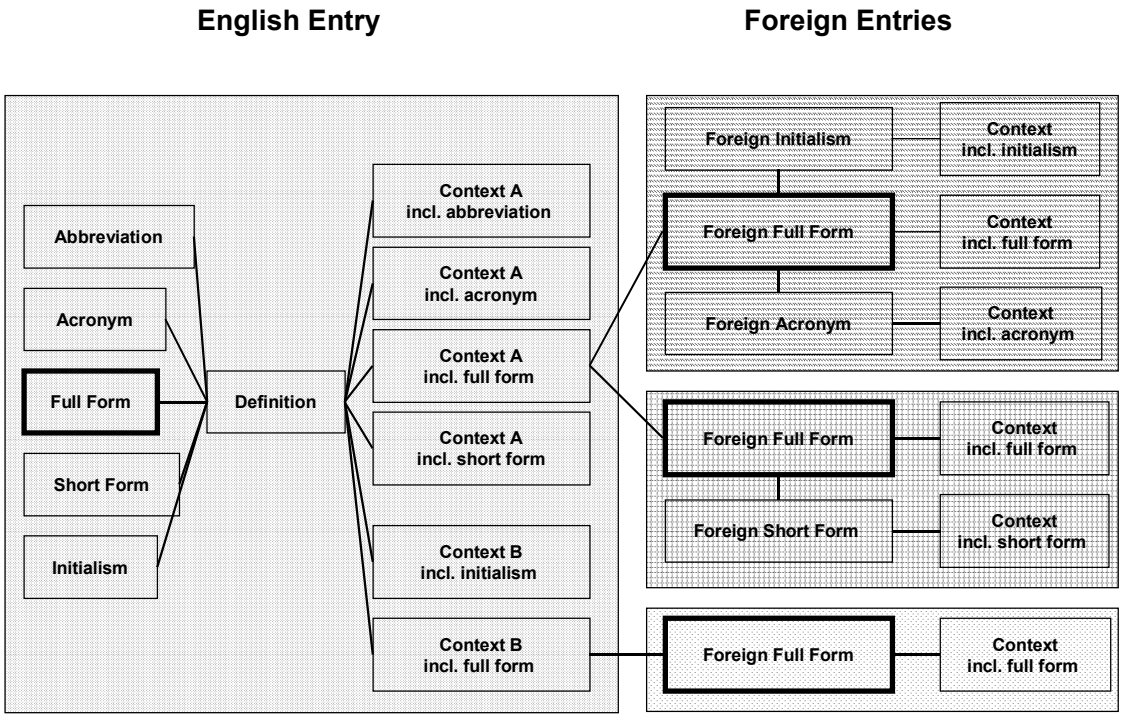

Figure 3: Links between designators, definition and contexts

First, there is only one entry per English concept, one of the greatest benefits of a conceptual database (Collet 2002). Furthermore, this English entry can contain as

many contexts as necessary to fully cover the English concept. Depending on these contexts, there may be several equivalents in one foreign language; in addition, if we consider the various representations a concept can have in English, the number of entries on the foreign side could increase even further. In this model, the relationship between representation, concept, context, and foreign equivalent as well as its context becomes much more evident.

This scheme presents obvious benefits for Asian languages, particularly Japanese and Korean. Term formation in these languages can occur in many different ways, and it is difficult, even for the experienced translator, to decide which representation of a concept is appropriate outside of a particular context.

By providing several English contexts, the characteristics of the concept will become more apparent to the English terminologist. In addition, research in the foreign languages will supplement missing information in English. For several months, we have been observing increased granularity in a particular entry as soon as translators and terminologists started researching the foreign term and providing feedback for the English entry. The problem of underspecification of a concept has been addressed in a forthcoming paper (Ceusters et al. *in press*). The authors proved that ontologies could be enriched through information gained by multilingual annotations.

There are also at least two risks inherent in this model. First of all, performance may be slow. We have not tested this proposal yet, but we are assuming that the links added between concept definition and foreign equivalent will slow down searches. More importantly, if the terminology database should become prescriptive in the future and we will compile a product-wide ontology, it will invariably indicate flawed concept designators. Whether the rework will be more extensive than the benefit of having multiple contexts remains to be seen.

8. Conclusion

The terminology look-up functionality tool will become accessible through the J.D. Edwards intranet. Using this look-up feature, consultants, marketing personnel, and software engineers will be able to search for any concept from anywhere in the world. The muscle client for translators and terminologists will most likely have several different interfaces serving different user groups.

While we are revamping the tool, more changes in process and in our linguistic standards will become apparent. Some of the changes have been described in this paper. In the next 12 months, we hope to satisfy the majority of the requests our terminology customers present to us.

References

Ceusters, W., I. Desimpel and S. Schulz. (in press) Using Cross-Lingual Information to Cope with Underspecification in Formal Ontologies. In R. Baud et al. (eds.). *Proceedings of MIE 2003*, St. Malo, May 2003.

Collet, T. 2002. Plaidoyer pour une théorie terminologique mettant sur pied d'égalité la fonction référentielle et la fonction discursive du terme. (Paper presented at *XVI World Congress of the International Federation of Translators*, Vancouver, Canada, August 2002.)

Karsch, B.I. and A.P. Pichler. 2000. If you have a question about a term, where do you go? (Presentation delivered at J.D. Edwards, Denver, U.S.A., June 2000.)

Karsch, B.I. and L. Searle. 2001. Terminology Management in Today's Fast-Paced Environment. (Presentation delivered at *STC Conference*, Portland, U.S.A., October 2001.)

Madsen, B.N., H.E. Thompsen and C. Vikner. 2002. Data modeling and conceptual modeling in the domain of terminology. (Paper presented at *TKE 2002*, Nancy, France, August 2002.)

Wright, S.E. and L.D. Wright. 2001. Terminology Management for Technical Translation. In S.E. Wright and G. Budin (eds.). *Handbook of Terminology Management*. Amsterdam: John Benjamins.

Multi-cultural and Multi-lingual Society: A Challenge for e-Health in South Africa

Nolwazi MBANANGA
The Medical Research Council, Pretoria, SA

Abstract: This paper reflects on cultural and linguistic diversity as challenges facing e-health and consumer health informatics in South Africa. The e-health concept is linked with another concept, that of consumer health informatics, in order to delineate the scope of the paper. e-Health and consumer health informatics and the relationship between these two concepts are defined. The challenges facing the provision of e-health and consumer health informatics in South Africa are explored in relation to cultural and linguistic issues. This is an exploration of steps towards use of information technology in providing health information and services. Culture in this paper is restrictive and not exhaustive. It refers to the culture of the consumers of health services as opposed to organisational culture that may include health services. Issues and questions that need to be explored further in the context of e-health in a South African environment are also discussed.

1. Background

South Africa has eleven official languages. Each of these is a main language which may be formed by several sub-languages. The diversity in cultures is characterised by subcultures, making South Africa a complex society. The concept 'society' in social theory has generally presupposed notions of cultural cohesion and social integration (Delanty 1998). It is hard to fit South African society within this definition. South Africa has a history of a government system that encouraged ethnic and tribal cultural cohesion and social disintegration. The system of divide and rule led to the complex diversities in cultures and languages experienced today. That South Africa one day became a democratic country did not mean that the ethnic segregation of the past disappeared the evening before. The segregation of the society left the country with a legacy of certain communities being underdeveloped, some developing and others developed depending on people's positions in relation to the old South Africa. The legacies of the old system in the country are reflected at every level of economic and health services development.

However, South Africa today is in the process of social and economic reconstruction and is harnessing every tool available, including technology, towards social integration. As an emerging society, South Africa is exploring the emerging information technology in resolving social issues including health services. e-Health is emerging as part of the solution.

2. e-Health and consumer health informatics

The concept 'e-Health' emerged in the mid-1990s, and refers to electronic health information and services available over networks such as the Internet and related

technologies (digital TV / WebTV, wireless media such as Web-compatible mobile phones and personal digital assistants; cf. Eysenbach et al. 2001). Consumer health informatics is an area of health informatics focusing on enabling consumers of health services and information to access valid and relevant information about their health status. Emergence of this area has been driven by factors such as evidence-based medicine, efforts to cut costs and promote health, and efforts to empower patients to help themselves and make informed choices. The paradigm shift towards realising the potential of patients and their families in health promotion and information technology also encourages an information age healthcare system (Eysenbach and Jadad 2001). Consumer health informatics involves analysing, formalising and modelling consumer preferences and information needs; developing methods to integrate these into information management in health promotion, clinical educational and research activities; investigating the effectiveness and efficiency of computerised information and network systems for consumers in relation to participation in health, health information development and health care-related activities; and most importantly, studying the effects of these systems on public health, patient health, professional relationships and society (Eysenbach and Jadad 2001; Mbananga *forthcoming*).

Consumers of health information must be able to access valid, relevant, appropriate and useful information about their health status and health risks. They should be able to judge the advantages and disadvantages of all possible courses of action according to their values, beliefs, preferences and personal circumstances such as level of socio-economic development (Eysenbach and Jadad 2001; Mbananga *forthcoming*).

The problem of low health literacy is in many cases due to poor understanding of health messages (Mbananga and Becker 2002). Members of the community with inadequate health literacy have complex difficulties related to communication, which may lead to poor health outcomes. This situation is compounded by much of the consumer information being produced at a higher reading level, visual images which are based on biomedical models and some areas using health terminologies which fail to communicate intended messages (Mbananga *forthcoming*).

3. Health information needs for consumers

The idea that people should be informed in order to help themselves, to make informed decisions and to promote their health and that of the communities, suggests the need for a service to meet health information needs. To be able to make informed decisions requires access to information about possible actions and courses to be taken and the advantages thereof. This demands that information should be packaged in a way that promotes integration of the individual's state of health, culture, norms, values and belief system (Eysenbach et al. 2001; Mbananga *forthcoming*). Eysenbach et al. (2001) suggest two different types of basic information required: patient-related information (diagnosis, pathology, personal risks), and general information about the external

medical evidence (effectiveness of different interventions for diseases). The role that can be played by information technology (e-health) in this regard is significant. How this can be achieved depends on fundamental issues such as: use of consumers' language in developing health information and services; the availability of health concepts and medical terms in all eleven official languages; the integration of different cultures in health care delivery and health information development; and access to information development. These are the challenges which are facing South Africa in terms of e-health.

4. Inverse information law

Hart explains the 'inverse care law' as the availability of good medical care which tends to vary inversely according to the need for it in the population being served. In analogy Eysenbach et al. (2001) postulates an 'inverse information law', i.e. access to information varies inversely with the need for it. Mbananga (*forthcoming*) reports that rural communities who are likely to have low health literacy levels do not have access to information because of poor dissemination networks, lack of infrastructure and poor access to information technology in South Africa. From all this one notices the following sequence: lack of health information (a consequence of inverse information law) can lead to poor health which can result in low income which leads to poor access to information technology to glean information. e-Health promises to bridge the gap between those who are information-poor and to reverse the inverse information law. However, there are a number of challenges to be overcome before e-health can be a solution in South Africa and elsewhere.

An important aspect of e-health services is to define what health knowledge or information is, because for the knowledge society the term 'knowledge' is different from what it could be regarded as by target groups (Mbananga *forthcoming*). The question is then what is knowledge, who should define it, and in whose metaphorical language?

The knowledge society, health professionals in this case, pose a number of new and old social problems and challenges. One wonders whether the information technology in health information dissemination and health services delivery is posing problems / challenges or creating opportunities. If the latter is the case, there is a need to investigate what kind of opportunities are created and how these will change how work has been done and services have been measured and consumed by target communities.

5. The Role of information technology in meeting e-health challenges

Computerised health information and services are provided to be assimilated by different communities and to make an impact on the health status of the community being served. The promises of the information age are that language technology will help in designing and implementing the systems that will solve the problems of

linguistic diversity. Speech recognition will help in interacting with a number of devices in people's languages. These systems will generate and present information in different languages through automated machine translators. Therefore, correct information will be delivered at the right time and in the language of the recipient. Editors will ensure that each version is culture specific. There is hope for a universal language at the horizon as a result of information technology (Ghonaimy 2000). The English language for now, dominates the Internet. In the case of South Africa it would appear that machine translation with detailed knowledge of languages and culture will be required at many levels.

6. Multi-lingual and multi-cultural information systems

Multi-electronic publishing will provide information to various audiences, helping them to select from options using information in respective languages (Ghonaimy 2000). The languages of different communities in South Africa are part of cultural diversity. Language has an intrinsic value closely connected with the cultural identity, knowledge and understanding. Information technology has a potential of shaping the future of cultures and languages. It is important to discuss how technology will affect the identity, culture and creative diversity amongst the communities in the country. Currently health information and services, both paper-based and electronic, are dominated by the English language, and medical terms and health concepts have been derived from English. The domination of English in health service environment creates divisions in health literacy (high and low health-literacy levels). A pertinent question is: Can culture and language diversity be preserved in health information development and health service delivery without losing the essence of these aspects? How much does it cost to do this preservation? These are the questions to be grappled with towards e-health in South Africa.

Within societies there are health professionals and theses are leading the way in the area of health service delivery and development. Their social positions, values and expectations differ fundamentally from the target group (Mbananga *forthcoming*). These health professional groups take a dominant position and define what knowledge mix is required for whom, what quality is good or bad, who is educated or not. The consequence is to overvalue professional knowledge versus the understanding of the importance of fundamentals and wisdom imbued in other information. There will be an increasing conflict between the minority of knowers and the majority who only understand living in traditional ways (Mbananga *forthcoming*). Likewise, Drucker (1994) suggests that poor communities with less knowers will turn into poor and ignorant countries. How this problem can be averted in South Africa centres on the issue of the development of health concepts and terminology in the eleven languages.

A significant step to be taken would be to find ways of preventing the clash between health professionals and consumers (Mbananga *forthcoming*). The clash is not only promoted by the difference between professional knowledge and consumer

knowledge only, but also by the lack of equivalent health concepts or terms in African languages in South Africa. The potential of a cultural clash might increase with the influx of different information. How to divert this cultural clash is going to be a challenge for information technology (e-health) in South Africa. There are notions that cultural information, through records, reports, maps and charts, discloses reality more widely than professional knowledge (Eysenbach et al. 2001). In situations where health records, and other general health information materials, do not include cultural information, reality is obscured. Socio-cultural information provides for reordering and enriching reality (Mbananga *forthcoming*). Information technology potentially lifts both the illumination and the transformation of reality to another level of power while it is also providing a new kind of information for easy access to consumers. Health information and services through the power of technology is supposed to be received as the arrival of reality. This can be realised if the language of consumers is used. Clearly electronic health information will need to be tailored to the socio-cultural and linguistic landscape of the South African population if it is to be useful and developmental. If this approach is neglected, information technology (e-health) will create a new division between haves and have-nots or deepen the old division.

7. Conclusion

There is a need to focus energies on devising uses for new technology that will enhance people's lives, improve health status and serve the values that people hold in common. How e-health will enhance the health status of ordinary people in South Africa depends on how the language and cultural diversity is incorporated in its conceptualisation and development. Central to this process is the development of health concepts and medical terms in the eleven languages of health consumers. This is posing a question: How far has South Africa moved towards the development of medical terminology and health concepts in African languages?

References

Delanty, G. 1998. Social Theory and European Transformation: Is there a European Society? *Sociological Research Online* 3/1. <http://www.socresonline.org.uk/3/1/1.html>

Drucker, P.F. 1994. The Age of Social Transformation. *The Atlantic Monthly* November 1994. <http://www.theatlantic.com/politics/ecbig/soctrans.htm>

Eysenbach, G. and T.L. Diepgen. 2001. The role of e-health and consumer health informatics for evidence-based patient choice in the 21st century. *Clin Dermatol* 19/1: 11-17.

Eysenbach, G. and A. Jadad. 2001. Evidence-based Patient Choice and Consumer Health Informatics in the Internet Age. *Journal of Medical Internet Research* 2/3. <http://www.jmir.org/2001/2/index.htm>

Eysenbach, G, C. Köhler, G. Yihune, K. Lampe, P. Cross and D. Brickley. 2001. A framework for improving the quality of health information on the world-wide-web and bettering public (e-)health: The MedCERTAIN approach. In V. Patel, R. Rogers and R. Haux (eds.). *Medinfo 2001, Proceedings of the Tenth World Congress on Medical Informatics*: 1450-1454. Amsterdam: IOS Press.

Ghonaimy, A. 2000. *Role of Language Engineering in Supporting Multilingual Aspects in Cyberspace*. <http://www.unesco.org/webworld/infoethics_2/eng/papers/paper_6.htm>

Mbananga, N. and P. Becker. 2002. Use of technology in reproductive health information designed for communities in South Africa. *Health Education Research Theory & Practice* 17/2: 195-209.

Mbananga, N. (forthcoming) *The Sociological Study of Reproductive Health Information in South Africa.* PhD thesis.

Terminology Coordination and Copyright Issues

Xolile T. MFAXA
National Language Service, Department of Arts & Culture, SA

Abstract: This paper explains the role of the *Terminology Coordination Section* (TCS) of the *National Language Service* (NLS), a chief directorate within the *Department of Arts and Culture* (DAC), in terms of its terminology management and coordination functions. It also concerns the challenges posed by copyright issues as far as terminology coordination and collaboration are concerned. The problems surrounding copyright emanate from the fact that TCS is the user of other people's work on the one hand, and the originator of terminological work on the other.

The functions of TCS are discussed with special emphasis on the collaboration of TCS with language structures throughout South Africa. The merits of collaboration are explained with special reference to the elimination of duplication as a cost-effective strategy of TCS. The issues surrounding copyright are discussed with the emphasis on TCS as both having a potential to infringe on others' copyrights and others infringing on its rights, especially as the originator and author of terminology products. The solution to the present copyright problems seems to lie with contractual agreements between the collaborating parties.

1. Terminology coordination

The national South African government has, through the *Department of Arts and Culture* (DAC), established an office in the *National Language Service* (NLS) for the coordination, management and modernisation of terminology in South Africa, the *Terminology Coordination Section* (TCS). TCS's main functions include facilitating the development and standardisation of terminologies in all the subject domains of all the official languages of South Africa. In this way TCS tries to move away from the notion that has been expressed by certain academics that the African languages are incapable of expressing abstract concepts (Cluver 1996: 6). More important, however, it seeks to coordinate all the terminological activities throughout the country by means of collaboration, establishment of partnerships and funding of terminological projects (Mfaxa 2001: 2).

Through these efforts duplication of similar terminology projects can be avoided or eliminated, as there is a central point where all the terminological projects and activities are registered. The aim of such a register is to make terminology users aware of lexicographic and terminographic services and projects currently being undertaken by national, provincial and local government, PanSALB structures, tertiary institutions, and individuals. Not only will the coordination efforts of TCS eliminate duplication but it will also promote the linguistic empowerment of all South Africans through terminological contributions that facilitate communication at different levels in various subject areas and domains of activity. The activities of TCS will ultimately lead to

structured and systematic distribution of terminological products in the form of technical dictionaries and term lists.

Apart from facilitating scientific and technical communication the TCS plays an important role in knowledge transfer, and consequently in the empowerment of South Africa's citizens through enhancing their technical and general linguistic capacity. In this connection, TCS has a structured annual programme whereby it selects collaborators and partners throughout the country for extensive training on terminology management principles and in use of technological tools for terminology management. The aim of this training is to establish and to reinforce partnerships with external professional collaborators with a view to enhance its production of standardised terminologies for the different subject fields and domains of specialised activity.

The NLS is particularly committed to devising plans and implementing strategies to speed up the process that would yield terminologies in the African languages and to fast-track the standardisation of such terminologies. Some of the plans include the following:

- Stakeholders should communicate with TCS before embarking on a terminology project. This goes a long way towards ensuring that there is no duplication of terminology projects.
- TCS encourages all stakeholders to undergo basic terminology and lexicography training before embarking on terminology work.
- All new projects regardless of whether they are undertaken by the TCS, the NLUs or the NLBs should be registered with TCS.
- All registered stakeholders have the right to access the term bank electronically via the Internet in order to provide terminology and/or to seek terminological information.
- TCS undertakes to identify different groups of organisations working in the same field or domain and to bring them together for collaboration and sharing of information.
- All collaborators undertake to actively encourage other organisations that are not registered with the TCS to do so in order to reap the benefits of being shareholders of the mainstream terminological process.
- Registration with TCS does not mean that the project belongs to NLS, it simply means that the project has received official acknowledgement and recognition by the government and permission will always be sought by NLS for utilisation of the project's terminology in any way.
- TCS has devised a comprehensive distribution and marketing plan for its products since 2002.
- The NLS will conduct regular needs assessment studies to ensure that relevant projects are being done with the limited human and financial resources available in South Africa.

The idea behind the TCS's endeavours is to exercise quality control, and to align and implement systems and structures to efficiently provide for the technical language needs of all language communities as well as to support the creation of integrated information networks. The rendering of facilities and skills to produce terminology and related products forms part of the spectrum.

This terminology coordination process will include the following:

- liaison between terminology stakeholders, role-players and collaborators (e.g. subject specialists and linguists);
- terminology development initiatives;
- ensuring the capturing of terminological data in the National Term Bank;
- strategies to disseminate terminological data.

One of the aims of TCS is to become a clearinghouse for terminology. The TCS would then gather and document multilingual and polythematic terminology in different registers in a variety of subject areas and domains. The terminology would become part of the National Term Bank and could be disseminated countrywide by means of terminology lists, technical dictionaries, CD-ROMs, and the *Human Language Technology* (HLT) virtual network.

2. Copyright issues and TCS

TCS, as has been stated already, develops its terminological products, not only by utilising the skills and expertise of the terminologists at its disposal. It also develops its terminology products with the assistance of collaborating groups throughout the country. Also, as was stated earlier, TCS has to disseminate terminological dictionaries and term lists to various role players like the *National Language Bodies* (NLBs), *National Lexicography Units* (NLUs), and other language structures in the country. The question that arises is how does the TCS protect its copyright. Another question also arises in as far as the copyright is concerned: How does TCS ensure that it does not infringe on the copyright of others when the terminologists are formulating definitions, using various sources of information. TCS makes its work available to the South African public at large in the hope that the South African copyright laws will protect its authorship against those that might infringe on its right as originators and authors of the terminology lists and technical dictionaries.

Copyright has been defined in many ways. For the purpose of this discussion, however, we are going to confine ourselves to the definition found in the *Oxford Advanced Learner's Dictionary of Current English* (Hornby 1989[4]: 63), where copyright refers to the '*exclusive legal right, held for a certain number of years, to print, publish, sell, broadcast, perform, film or record an original work or any part of it'*. In the South African context, for example, the number of years under which a particular work is protected is 50 years. After fifty years, the work is no longer under

the protection of copyright. Alberts & Jooste (1998: 124) maintain that copyright law technically only extends as far as the law permits it to, and as far as the geographical boundary permits the law to rule. This means, for example, that copyright laws in South Africa are unique to, and are applicable only in, South Africa. However, the existence of electronic communication networks and database storage systems across countries have now complicated the issue of copyright. These modern systems have prompted terminologists to think globally in terms of copyright issues, so that authors are protected across national borders. The Berne Convention makes provision for the protection of authors' copyright across nations. This convention protects the works of translators, literary writers, and artists, among others.

Galinski & Wright (2001: 297) argue that in some countries, when salaried employees of a particular organisation have created terminology, and these employees performed this function as part of the conditions of their employment, this then means that the employer has copyright of their work. Taking the case of TCS terminologists who compile terminology lists for the government, the answer becomes clear – the copyright of the work they perform belongs to the government, in particular, the Department of Arts and Culture, since they are paid for the work they do. Galinski & Wright (2001: 279) further state that the contracting or authorising parties, who have defined the resulting body of data as part of the scope of work, have the copyright. Thus TCS, as the one contracting the collaborators all over South Africa, also has the copyright of the work performed by the terminology collaborators who are developing terminology on its behalf through contracts.

TCS is not only concerned with infringement of its copyright as the originator and author of terminology work, it is also concerned about infringing on the copyright of other originators and authors of works. For example, TCS terminologists excerpt their terms from written authoritative sources and even the definitions they use in the documentation of their terminology have been sought from written sources. It is therefore clear that a terminological entry usually contains at least one, and often numerous, references taken from published, proprietary, or standardised works. Thus in a sense TCS could as well infringe on the copyright of others. The terminologist seems to be safe in his/her role as the excerpter of terminology since Galinski & Wright (2001: 287) state that the existing judicial decisions and precedents on copyright as it affects lexical content clearly do not view the names, facts, lists of observations, and words and idioms as worthy or capable of being protected as units of intellectual property. Galinski & Wright further state that units smaller than a complete sentence cannot be construed as copyrightable.

The argument in the foregoing two paragraphs is that a terminologist has a dual relationship with copyrightable material. He/she is both the consumer and provider of potentially copyrightable material. The terminologist or the employer of the terminologist thus needs protection as much as others need protection from him/her. Alberts & Jooste (1998: 128) say that infringement of copyright occurs when a person

commits an act that is the sole prerogative of the copyright holder without the permission of the copyright holder. One is regarded as having infringed on the copyright of another when one has translated, reproduced, published, performed, broadcasted, or adapted literary work in any manner or form, without the consent of the copyright holder.

3. Conclusion

This overview has highlighted the work of TCS as far as collaboration and terminology coordination are concerned. It also focused on some of the collaboration problems that TCS is likely to encounter as far as copyright issues are concerned. The questions surrounding the ownership of copyright in contract work, the excerptions of terminology, definitions and translation work, have been dealt with.

The best way in which TCS can protect its copyright as far as the database is concerned, is to enter into agreements with its collaborators. Galinski & Wright (2001: 300) maintain that such contractual agreements can support good faith collaboration on the part of partners who might otherwise be unwilling to share their data.

References

Alberts, M. and M. Jooste. 1998. Lexicography, Terminography and Copyright. *Lexikos* 8: 122-139.

Cluver, A.D. de V. 1998. Language Development. In *Proceedings of a LANGTAG Workshop, March 1996*. DACST: Pretoria.

Galinski, C. and S.E. Wright. 2001. Intellectual Property Rights. Copyright and Terminology. In S.E. Wright and G. Budin (eds.). *Handbook of Terminology Management*. Amsterdam: John Benjamins.

Hornby, A.S. 1989[4]. *Oxford Advanced Learner's Dictionary of Current English*. Oxford: OUP.

Mfaxa, X.T. 2001. *Guidelines for a sound working relationship between the Terminology Coordination Division of the National Language Service (NLS) and the stakeholders in the Provinces*. DACST: Pretoria.

South African Copyright Act 98 of 1978 (as amended). Pretoria.

The Role of Terminology and Classifications for Knowledge Bases

Sergey PAPAEV
The Russian Research Institute for Classification, Terminology and Information on Standardization and Quality (VNIIKI), Russia

All the problems of linguistics referring to the clarification of a term concept bring it nearer not only to different fields of scientific knowledge but to various spheres of industrial and professional activity as well. It is here where we may observe the liaison between the development of language, its lexical system and the history of material and spiritual culture of a nation. The history of terminology in certain branches of science, culture and industry may be regarded as a kind of story dealing with acquisition of knowledge of nature and society.

Terminology development should be considered not only as a national-historical problem but also as an international problem, namely the problem of world science and human civilizations, and more in general, the history of cultural cooperation between people.

The present period of human community development may be characterised as the period of informatisation. Information is becoming the main resource which practically determines the possibilities of further development. During the period of information production and consumption it is necessary first of all that the society continues to function, that the national economy is managed, and that comfortable conditions for human life and activity are created.

Informatisation principally requires the development of new technologies and special intellectual instruments. In a market economy, information appears to be both a kind of commodity of great importance and a product of a specific nature, that means information may be consumed without being reduced and be quickly delivered at large distances.

It is useless to speak of informatisation without terminology that is supposed to be a special kind of intellectual product. This product is widely used for the development and operation of new information technologies and intellectual systems used within the framework of knowledge engineering.

However, if we consider terminology as an intellectual product, then it is necessary to discuss the quality and reliability of this product as well as its special properties which

permit it “to meet the requirements”, “to be functionally fit”. The nomenclature of this intellectual product should also be considered.

It should be mentioned that the following problems of terminological science become urgent: to find the optimal way to include terminological products in the general theory of quality and reliability, to determine its place in the technological policy both of a particular state and of the whole world community, to determine the optimal proportion between the free initiative in the process of its production and the special control and guidance of this process. Surely questions of costs of its production should be considered on a regular scientific basis as well.

As an independent discipline artificial intelligence (AI) has a rather short but interesting history of nearly thirty years. The term itself serves as a name for a rather large family of intellectual systems and theoretical researches in the field, theoretical and practical aspects of terminology applications being investigated in a number of them. But we mean terminology of a rather high level and quality and first of all standardised and ordered terminology appears to be of this kind.

It is reasonable to delimit two groups of terminological products each of which has its own specificity, fulfils different functions, and on the whole they may be used in the sphere of knowledge bases.

We believe that terminology included in standards and vocabularies may be referred to the first group. Terminology represented in them is an ordered, normalised set of terms free from errors and faults which are only natural for a particular sphere and reflect the corresponding system of concepts belonging to a certain field of activity.

In other words, standardised terminology in standards and vocabularies contains compact descriptions of essential characteristics and relations that a certain object of human activity possesses. Information presented in the system of definitions with the help of which terms are introduced in the documents mentioned above permits not only to retain the knowledge gained by using certain procedures but to get new information, new knowledge, as well. For instance, if the standard comprises such concepts as “risk assessment”, “risk estimation” and “risk evaluation”, then using operational definitions we can get information by means of distinguishing between these terms.

Terminology used to construct denominations in the classifications of technical and economic information may be attributed to the second group. These two kinds of terminological products belonging to the first and second groups have a lot in common because in both cases the principle of systematic hierarchical construction of material is used. As a rule, denominations used in classifications of products, processes and

services are constructed on the basis of standardised terms. In many cases the structures of the classification and of the system of terms of a standard (or standards) may coincide to a great extent.

However one should remember that the main function of the classification is to represent a certain nomenclature of products, processes or services (often arranged in accordance with its characteristics of generalisation). Classifications are also directly linked to factual data concerning a particular nomenclature.

Nowadays, within the framework of the Federal Program 28, Russian classifications are being developed and improved on the basis of former USSR classifications, covering main types of technical, economic and social information used for inter-branch exchange, such as types of economic activities, products and services, units of measurement, currencies, etc. Basic requirements for classifications are the orientation to the market economy and the harmonisation of Russian classifications with international and regional ones (UN, EC, ILO, ISO).

For example, the Russian Classification of Standards is harmonised with the International Classification of Standards (ICS). From our point of view the adoption of the ISO International Classification of Standards is a positive step. Its implementation has considerable influence not only in the field of classification but in the sphere of terminology standardisation and harmonisation of the international information retrieval languages as well. Development of the Russian version of the International Classification of Standards proved once more that it was necessary to use standardised terminology in the document, hence uniform and unambiguous terminology. Most difficulties we found while developing the Russian version were due to an insufficient degree of harmonisation of the Russian and English terminology. It should be noted that there is not enough information on results of the work in these fields of terminological activity. Coordination of activity in the sphere of terminology and classifications development is urgently needed, especially at regional and international levels.

As a result of the systematised terminology of special languages contained in the terminological databank “Rosterm” and in the database of Russian classifications functioning in VNIIKI, there are favourable conditions to use ordered terminology to develop terminological support for intellectual systems that might be created in various branches of the national economy. Usage of both standardised terminology and denominations of classifications increases the potential capabilities of various intellectual systems, such as expert systems, training systems, and others. In this case both intension and extension approaches to information seem to be combined, permitting to take advantage of acquiring new information (new knowledge).

Both groups of terminological products should have the following common characteristics:

- the vocabulary represented in these groups should have standardised and ordered terminology as its basis;
- the terminology used in them should be either optimally harmonised or compatible;
- standardised and harmonised terminology should be of a quality and quantity that provides knowledge transfer and access to information in a certain field.

The development of terminological documents in the informatisation and utilisation of artificial intelligence systems or intellectual systems should stipulate the possibility of conversion of terminological information into its formal representation that would use some kind of knowledge representation language, for instance a frame representation language, a semantic network representation or a predicate logic. In other words, nowadays while determining the structure and number of terminological standards, planning terminological work both within the framework of a particular organisation and at national or international levels, choosing characteristics to introduce them into definitions of the terms and formulating definitions, one should take into account whether the systems of terms suggested may be easily and adequately represented in a formal language of knowledge representation, and see which terminology is mostly needed for intellectual systems.

We believe that this apprehension of terminology as an intellectual product will be predominant, and this product will be in great demand. Today one of the most prospective branches of terminological scientific research is analyses of scientific, technical and other kinds of texts from the point of view of utilisation of terminology contained in those texts; namely usage of terms and their definitions for knowledge representation, for gaining new knowledge.

Today it is quite clear that terminological databanks may and should be used to develop knowledge bases. This is certainly true for databanks that contain standardised and recommended terminology. A terminological standard (or recommendation) that reflects a concept system of a certain field of science and technology represents some kind of structured knowledge in the field. Besides, standardised and recommended terminology may be regarded as a dynamic lexical stratum which develops and changes together with development of science and technology in the process of knowledge acquisition. In practice it is implemented by way of either introducing amendments and alterations in the standards, or by way of their periodical revision. It should be noted that terminological standards are not isolated from each other, they are interlinked. So the standardised terminology may in general be regarded as a large, complex, structured system; its subsystems and components are represented by

standardised terms and their sets, namely by terminological standards. This is an important fact, because particular emphasis should be placed on the problem of establishing relationships between the concepts and choosing classification schemes to structure knowledge while developing knowledge bases.

The composition of data element sets used to describe terminological entries is also of considerable interest. Terminological databanks where the wide range of elements is represented in data element sets are preferred to serve the purpose of a knowledge base development. The following data elements are the most important ones: definition of the concept, notes and explanations for the definition, context, examples of term usage – or in other words the information that permits to acquire knowledge about the real thing the term denominates. Data elements that establish systematic relationships between concepts are also significant (for instance gender and specific difference, whole and part, etc.). In conclusion it would be reasonable to name the following directions of terminological activity which we believe are important for artificial intelligence systems:

- Determination to what extent systematised terminological standards (dictionaries) are provided for those spheres of activity where there is a high potential need for development of artificial intelligence systems. Updating of standardised terminology that would take into account requirements of artificial intelligence systems. Updating means introducing amendments, additions, and the development of new standards.
- Determination of the groups of artificial intelligence system users and of their needs as well as of those fragments of terminological systems that might satisfy these needs best.
- Development of the main principles and methods of use of terminological databases for the construction of knowledge bases.
- Determination of the set of requirements and recommendations in the field of terminology standardisation and ordering that would create the necessary prerequisites for recording information of terminological standards and dictionaries in a formal form employing various languages of knowledge representation.
- Development of an ordered and harmonised terminological system in the field of artificial intelligence, that would ensure unambiguous and non-contradictory understanding of scientific and technical documentation in the sphere of artificial intelligence.
- Development of a training terminological system within the framework of InfoTerm, together with the help of involved national organisations would permit, on the one hand, to make the process of terminological training more active and extensive and, on the other hand, to give experts in terminology abundant material to solve problems of terminology use in the field of artificial intelligence.

Special Multiobjective and Multilingual Knowledge of Electronic Encyclopaedia

Rodmonga K. POTAPOVA[°] & V.V. POTAPOV[‡]
Department of Applied and Experimental Linguistics, Moscow State Linguistic University, Russia[°] &
Department of Philology, Moscow State Lomonosov University, Russia[‡]

Abstract: This paper presents the main principles and overall structure of a linguistic database of an *Electronic Encyclopaedia for Russian* (EER). The EER is intended for several kinds of users (e.g. for researchers, teachers and students in the field of Russian; for forensic phonetics experts; as reference expert system in the domain of theoretical and practical knowledge of Russian; etc.).

1. Introduction

The *Electronic Encyclopaedia for Russian* (EER) is developed on the basis of the following information:

- an electronic knowledge database in the domain of the linguistic sciences and Russian;
- methods of practical activity of linguistic experts, especially for forensic phonetics purposes.

The linguistic part of the EER includes three kinds of databases (three knowledge blocks):

- textual definitions which contain the main linguistic knowledge in the field of language and speech theory;
- a dictionary of special terminology on linguistics, speechology and forensic phonetics;
- a collection of bibliographic references.

This encyclopaedia may be used at the same time as a detailed electronic handbook on language and speech of Russian and as an expert system in forensic phonetics. More accurately, it is hard- and software support used by linguistic corpora of specialists and forensic phonetic experts which will use the linguistic knowledge for several professional purposes.

Before describing the structure of the EER we need to explain a problem of appearance of this electronic product. The first aim was to prepare and to create an electronic information support for forensic phonetic experts. It is known that experts in the process of identification of a speaker must examine the voice and speech of each speaker with respect to aspects like:

- sex;

- age;
- voice quality;
- language / dialect / variety;
- prosody (intonation, rhythm, stress, etc.);
- pronunciation of single speech sounds;
- paralinguistic features;
- extralinguistic features.

2. The structure of the EER

The linguistic information (lexical, phonetic, syntactic, semantic and other features) lies at the basis of expert knowledge. The same knowledge would be necessary for researchers, teachers and students in the field of Russian.

The structure of the EER is as follows: the first block of the linguistic database of the EER contains a set of semantic fields with the following keywords:

A. Fundamentals of the main concepts "language" and "speech".
B. Speech production.
C. Speech perception.
D. Multilevel linguistic information.
E. Paralinguistic information.
F. Extralinguistic information.

Every semantic field includes a set of sub-fields with textual definitions of every sub-keyword. The structure of these semantic fields can be illustrated on the basis of semantic Field A as follows:

I.0. Language and speech
I.0.1. Natural language / artificial language
I.0.1.1. Natural language: native / non-native one
I.0.1.2. Standard language: norm or norms (for Russian)
I.0.1.3. Relations between standard language and dialects, sociolects, etc. (for Russian)
I.0.1.4. Natural and artificial bilingualism, multilingualism, etc. (for Russian and other languages)
I.0.1.5. Linguistic interference (for Russian and other languages)
I.0.1.6. Natural language and other semiotic systems of human communication
I.0.2. Speech: writing / speaking
I.0.2.1. Spoken language (speech) communication (arts of speech communication models)
I.0.2.2. Speech behaviour, speech activity
I.0.2.2.1. Kinds of speech activities: reading (free, half-free), speaking (free, half-free)

I.0.2.2.2. Varieties of speech communication: monologue, dialogue, polylogue
I.0.2.2.3. Differentiation of concepts: speech act, speech activity, speech material (corpus)

The EER contains more than 250 sub-keywords, all with textual definitions and linguistic examples.

The special parts of the linguistic block includes phoneticised orthography (phonetic transcription). We use a style of transcription which is current in Russian and common among Slavists on the European continent. In any phonetic transcription a more or less arbitrary choice has to be made concerning the degree of delicacy. A less delicate, broad transcription shows fewer phonemic (phonologic) distinctions. A more delicate, narrow transcription shows more phonetic niceties. In the EER we used the second type of transcription for segmental and suprasegmental description of spoken utterances (e.g. special signs of primary and secondary stress, nasalization, length, syllable division, united pronunciation of two consonants, sound and phrase borders, hardness / softness of consonants, stress, voicing / invoicing of vowels, etc.

The terminological block of the linguistic database of the EER (cf. Potapova 1997) includes 300 lexical items in alphabetical order and can be increased. We may plan to add a corpus of world languages and their linguistic characteristics (cf. Potapov 1997).

All blocks of the EER are connected by means of hypertext technology and designed as Help-files, possessing all the properties and advantages of Windows WinHelp systems. The first version of the EER was described without linguistic details in Potapova (1999) and Potapova & Potapov (1999).

3. Conclusion

This paper described for the first time the detailed structure of the linguistic database of the EER and new steps in the evolution of this electronic product (the EER, version 2). The linguistic part (all three blocks of the database) handles multilevel linguistic information, terminological peculiarities and bibliographic sources of modern Russian.

The previous version of the EER can be characterised as a multimedia information system with audio and video indexing (various lexical items in spoken Russian, their morphologic sub-items, spoken texts; acoustic waveforms of the spoken material; visual support of segmentation rules of all kinds of spoken language).

The new version of the EER is based on an integration of diverse kinds of knowledge about language, spoken language (speech) and applied domains. This version shows new possibilities to integrate different knowledge sources (theoretical, experimental, perceptual, acoustic, etc.). We plan to continue the elaboration of the EER and to translate the Russian text into other languages.

References

Potapov, V.V. 1997. *Brief linguistic reference book: languages and scripts*. Moscow.

Potapova, R.K. 1997. [in Russian] *Speech: communication, information, cybernetics*. Moscow.

Potapova, R.K. 1999. Some aspects of forensic phonetics expert learning (on the basis of Russian). *SPECOM'99*. Moscow.

Potapova, R.K. and V.V. Potapov. 1999. Database of forensic phonetics knowledge (as applied to an electronic encyclopaedia for Russian experts). *Proceedings of the Annual Conference of IAFP*. York.

Appendix

Some extracts from the EER are presented below. These articles are samples from the multilingual Russian-English-German-French database for new information technology, which includes definitions and context examples.

A08 D: Adresse (f) ~, ~n
E: address
F: adresse (f)
R: адрес
Die Ad. ist die Kennzeichnung eines Speicherplatzes im Arbeitsspeicher eines Computers, bzw. eines Massenspeichers. Mit Hilfe einer Ad. können der Inhalt eines Speicherplatzes gefunden bzw. Daten dort abgelegt werden.

B16 D: Bit (n) ~s, ~s
E: bit
F: bit (m)
R: бит
Das Bit, zusammengesetzt aus binary and digit, ist eine binäre Informationseinheit. Zur Kennzeichnung der Zahl der Speicherzellen (Speicherkapazität) wird der Zusatz 1K=2^10=1024 verwendet. M (Mega) und G (Giga) stehen, beziehungsweise, für 10^6 und 10^9 als auch für 2^20=1024K und 2^30=1024M.

C05 D: Chip (m) ~s, ~s
E: chip
F: puce (f), pastille (f)
R: кристалл, чип, микросхема
Bauteil, das aus einem einkristallinen Halbleiterplättchen (im allgemeinen Silizium) von einigen mm2 bis zu cm2 (typisch 10mm2 bis 60 mm2) Fläche besteht, auf das mit dem Verfahren der Halbleitertechnik Strukturen zur Realisierung integrierter Schaltkreise aufgebracht sind.

D09 D: Datenbanksystem (n) ~s, ~e
E: database system, databank system
F: systeme (m) de banque de données
R: система банка данных
Sammelbegriff für die Datenbank und das zugehörige Datenbankverwaltungssystem.

E20 D: Expertensystem (n) ~s, ~e
E: expert system
F: systeme (m) expert
R: экспертная система
Anwendungsprogramm aus dem Bereich der künstlichen Intelligenz, das ähnlich wie ein menschlicher Experte auf einem bestimmten Fachgebiet bei der Problemlösung und Beratung behilflich ist. Ein Es. besteht im Wesentlichen aus einer erweiterbaren Wissensdatenbank und einem Inferenzsystem (Schlussfolgerungssystem). Expertensysteme werden z.B. bei der medizinischen Diagnose eingesetzt.

F19 D: Fraktale (npl)
E: fractals
F: fractales (fpl)
R: фрактали
Durch komplexe mathematische Berechnungen entstandenes graphisches Gebilde, das immer ein ähnliches Aussehen besitzt, unabhängig davon, wie stark dieses vergrössert wird. Viele Gebilde in der Natur, z.B. Pflanzen, gehören fraktalen Gesetzen. Heute werden Fraktale in vielen Bereichen genutzt, z.B. in Kinofilmen zur Erzeugung realistisch wirkender Landschaften und zur Datenkomprimierung (fraktale Bildkomprimierung).

G11 D: Graphiksystem (n) ~s, ~e
E: graphics system
F: systeme (m) graphique
R: графическая система
Menge graphischer Manipulationsfunktionen, mit denen einem rechnerinternen Modell ein Bild erstellt und auf graphischen Ausgabegeräten dargestellt werden kann und umgekehrt, veränderte oder neu erstellte Bilder von den graphischen Eingabegeräten in ein rechnerinternes Modell überführt werden kann.

H15 D: Hypertext (m) ~es, ~e
E: hypertext
F: hypertexte (m)
R: гипертекст
Prinzip bei der Bildschirmdarstellung, die es erlaubt, Textstellen durch Blättern einzusehen und dem Leser die Wahlmöglichkeit bietet, welchen Text er als nächsten lesen will. Ht. kann nur gelesen werden. Mit Ht. lassen sich

Dokumente verknüpfen. Statt zu zitieren, genügt z.B. ein Verweis (hyperlink oder hyperword) auf das ganze zitierte Original-Dokument, das auf Wunsch des Lesers aufgerufen und an der zitierten Stelle gezeigt wird. Ausserdem besteht keine Beschränkung auf herkömmlichen Text, vielmehr können auch Bilder, Töne oder Filme eingebunden sein.

I12 D: Internet-Protokoll (IP) (n) ~s, ~e
E: Internet protocol (IP)
F: Internet protocole (m)
R: интернет протокол
Das IP ist ein Übertragungsprotokoll für Software, die die Internet-Adresse ermittelt und Nachrichten versendet bzw. empfängt.

J01 D: Java-Sprache (f) ~, ~n
E: language Java
F: langage (m) Java
R: язык Ява
Programmiersprache für WWW-Browser von Sun. Java-Programme laufen in sogenannten virtuellen Maschinen ab, die vom Browser unterstützt werden müssen. Dadurch läuft Java unabhängig vom verwendeten Betriebssystem.

K03 D: Koaxialkabel (m) ~s, ~
E: coaxial cable
F: cable (m) coaxial
R: коаксиальный кабель
Kabel für hohe Datenübertragungsraten. Ein Kabel besteht aus dem Innenleiter (transportiert das Datensignal) einer nichtleitenden Schicht (Dielektrikum), einer elektrischen Abschirmung (Aussenleiter, Metallgeflecht) und einer Aussenisolierung.

L13 D: Lumineszenzdiode (f) ~, ~n; lichtemittierende Diode (f) ~, ~n
E: light emitting diode (LED)
F: photodiode (f)
R: фотодиод
Halbleiterelement, das bei geringer elektrischer Leistung (z.B. 10mA und 2V) und hoher Lebensdauer Licht ausstrahlt. Es gibt (infra-rote) IR-Dioden, und rote, grüne, gelbe, sowie blaue LEDs. Sie werden heute in fast allen elektronischen Geräten zur Informationsanzeige benutzt.

M11 D: Maschennetz-Topologie (f) ~, ~e
E: meshed topology
F: topologie (f) maillée
R: топология типа сетка
Netzwerktopologie, bei der jede Arbeitsstation über ein separates Kabel mit jeder anderen Arbeitsstation verbunden ist. Der Vorteil liegt in der hohen

Ausfallsicherheit, der Nachteil in den hohen Verkabelungsaufwand, dessentwegen diese Topologie nur selten verwendet wird.

N01 D: Nadeldrucker (m) ~s, ~
E: needle printer
F: imprimante (f) par points, imprimante (f) a matrices d'aiguilles
R: печатающее устройство с однорядным (игольчатым) знакосинтезирующим механизмом, одноряное (игодьчатое) знакоситезирующее устройство, игодчтое печатающее устройство
Drucker, der über einen mit (9 oder 24) Nadeln bestückten Druckkopf verfügt, der auf einer Schiene bewegt werden kann. Zwischen Papier und Druckkopf ist ein Farbband gespannt, gegen das die einzelnen Nadeln des Druckkopfes gedrückt werden und dadurch auf dem Papier Punkte erzeugen. Nadeldrucker sind langsam und sehr laut, jedoch billig in der Anschaffung und Wartung.

O07 D: Optische Zeichenerkennung (f) ~, ~en
E: optical character recognition (OCR)
F: reconnaissance (f) optique des caracteres
R: оптическое распознавание символов
Verfahren zur Erkennung von Texten, die mit einem Abtaster eingelesen wurden. Dabei wird die durch den Einlesevorgang erzeugte Bitmap-Graphic in Textinformationen umgewandelt. Bei der Texterkennung unterscheidet man zwei Verfahren, Mustervergleich (pattern matching) und Characteristikenerkennung (feature recognition).

P12 D: Plattform (f) ~, ~e; Rechnerplattform (f) ~, ~e
E: platform
F: plateforme (f), plate-forme (f)
R: платформа
Überbegriff für eine Betriebsumgebung für Programme Bei einer Pf..kann es sich sowohl um ein spezielles Computersystem als auch um ein Betriebssystemhandeln. Viele Anwendungsprogramme werden für mehrere Rechnerplattformen angeboten.

R07 D: rechnergestützte Entwicklung (f) ~, ~en; rechnergestützte Ingenieurtütigkeit (f) ~, ~en
E: computer-aided engineering
F: ingénierie (f) assistée par ordinateur
R: автоматизированная разработка, автоматизированное моделирование
Überbegriff für die Bereiche computergestützter Entwurf (CAD), computergestützte Entwicklung (CAM) und computergestützte Planung (CAP).

Towards Second-Generation Spellcheckers for the South African Languages

D.J. PRINSLOO[°] & Gilles-Maurice DE SCHRYVER[‡]
Department of African Languages, University of Pretoria, SA[°‡] &
Department of African Languages and Cultures, Ghent University, Belgium[‡]

Abstract: In this paper we present spellcheckers for the South African languages, viz. for the nine official African languages and for Afrikaans (English already being catered for in the group of world-English spellcheckers). The first section is devoted to (i) describing certain basic aspects regarding the functionalities of spellcheckers, and to (ii) some specific African-language issues. This is followed by a brief evaluation of spellcheckers currently available for Afrikaans and some of the African languages. The final part deals with more advanced principles underlying spellcheckers with a view to create the next generation of spellcheckers for the South African languages.

1. Human Language Technology (HLT) and spellcheckers

From the early 1960s onwards, researchers have designed various methods for the automatic detection of erroneous words in running text. Today, four decades later, there isn't any self-respecting word processor that doesn't include a spelling checker, as well as a spelling suggestor and/or corrector, a grammar checker and even a thesaurus as an integral part. This is true for all languages with significant worldwide commercial importance, less so for those languages with a limited commercial value. When we focus on the African languages[1], we must sadly note that commercially available spellcheckers are unfortunately the exception rather than the rule. It can hardly be disputed that the use and development of spellcheckers for the African languages at large are still in their infancy. For most African languages no spellcheckers exist and for those languages for which spellcheckers are available, the actual use is questionable.

All efforts regarding state-of-the-art, high-tech development of especially the African languages should be applauded. We believe however that such activities and the development strategies should be sensitive to certain realities of the South African situation and should address Human Language Technology (HLT) needs on a *priority basis* rather than on an *ideal*-HLT-development schedule. This means that major projects should be designed in such a way as to render *regular spin-offs*, i.e. usable products that are urgently needed. This might even entail taking shortcuts in the short term in order to provide products for immediate use for which the technology is in real

[1] Since this paper is being submitted for publication in South Africa, necessary sensitivity with regard to the term 'Bantu' languages is exercised in the authors' choice rather to use the term *African* languages. Keep in mind, however, that the latter includes more than just the 'Bantu Language Family'.

terms still under development. African languages in particular need what we call first-generation spellcheckers *now* to satisfy the immediate needs which could be described as spellcheckers that can detect most incorrectly typed words and suggest alternatives. This should be followed by subsequent, more sophisticated and improved spellcheckers which can also check grammatical structures. We thus believe that if ways can be found to satisfy the immediate needs of the users of specific languages, the process should not be delayed simply for the sake of releasing a more sophisticated spellchecker as the first product.

2. Brief theoretical conspectus on spellcheckers

The term 'spellchecker' is used here to cover what the average user understands under this term today, i.e. a piece of software, mostly integrated into a word processor like Microsoft Word or Corel WordPerfect, which (i) *checks* for spelling (and grammatical) errors, (ii) *automatically corrects* some typos, (iii) makes *suggestions* for other mistakes, and (iv) often includes a *thesaurus* (i.e. a list with synonyms and antonyms).

Viewed from the angle of the compiler of a spellchecker, Kukich (1992), still one of the definitive reference works, points out that three types of distinctions must be made: (i) error *detection* versus error *correction*; (ii) *interactive* spelling checkers versus *automatic* correction; and (iii) attention to *isolated* words versus linguistic or textual *context*. These distinctions result in the fact that research in this field *'has focused on three progressively more difficult problems: (1) non-word error detection; (2) isolated-word error correction; and (3) context-dependent word correction'* (Kukich 1992: 377).

Basically there are two main approaches to spellcheckers. Firstly, one can program software with a proper description of a language, including detailed morphophonological and syntactic rules, which computes over a stored list of word-roots. Secondly, one can simply compare the spelling of typed (or scanned) words with a stored list of top-frequency orthographic word-forms.

3. Issues in the design of spellcheckers for the South African languages

As far as we know, the only commercially available spellcheckers for the African languages are the one developed for Kiswahili by Arvi Hurskainen (Microsoft Word; cf. Hurskainen 1999: 139), and the first-generation series for isiZulu, isiXhosa, Sepedi and Setswana developed by D.J. Prinsloo (Corel WordPerfect; cf. Prinsloo & De Schryver 2001: 129). Spellcheckers known to the authors for Afrikaans are the commercially available products by Corel WordPerfect, Pharos, and the University of Potchefstroom for CHE.

In oversimplified terms it can be said that the purpose of a spellchecker in word processing software is to alert the user to possibly incorrectly-typed words or strings and to suggest options for correction. It can of course be argued that the principles underlying error detection and the techniques to suggest improvements are language-

independent. There are however certain unique characteristics of African languages that require adjustments in the approach to e.g. error detection. A good example in this regard is the handling of occurrences of sequences of equal words. One of the typical errors made in text production in any language is indeed the erroneous repetition of a word (*the the* is common in English). Therefore, a standard error-detecting function in spellcheckers is to highlight occurrences of supposedly-erroneous sequences of equal words. For the disjunctively-written African languages this, unfortunately, results in the highlighting of a huge number of *correctly typed double, triple, ... words*. For these languages this function is counterproductive because it delays the process of verification of correctness rather than contributing to it. Secondly, the handling of special characters in spellcheckers for African languages is a problematic issue. Ideally, provision should be made for all 'special characters' (i.e. those with a Latin base cum diacritics) used in these languages such as *š* and *Š* in Sepedi, and a fairly extensive number for Tshivenda, just as is the case for special characters like *ø* in Danish or *ç* in French. The Sepedi *š* and *Š* pose no problem for either compiler or user of spellcheckers since these characters have been assigned standard ASCII values, namely 0154 for *š* and 0138 for *Š*. Both programmer and user can therefore easily create them. This, however, is not the case in Tshivenda where the average user does not have a special character set on his/her computer. Moreover, albeit words typed without the diacritics could even be (semi-)automatically converted to the correct orthography by a Tshivenda spellchecker, such texts will create problems in printouts, e-mail correspondence and Internet up- or downloads and this will in the end be counterproductive unless certain specific solutions could be found.

4. A brief evaluation of currently available spellcheckers for the South African languages

In this section answers are sought to the questions:

- Is it possible to obtain acceptable error-detection levels for South African languages using spellcheckers solely based on top-frequency wordlists?
- What does the average user regard as a minimum or satisfactory level of success?
- Will the success rate be comparable for conjunctively and disjunctively written languages? Or thus, should a different approach be followed for the Nguni languages (isiZulu, isiXhosa, siSwati and isiNdebele) on the one hand, and the Sotho languages (Sepedi, Sesotho, Setswana) as well as Tshivenda and Xitsonga on the other?

A statistical evaluation – part of a much larger study – of the situation for Afrikaans, and then for isiZulu and Sepedi will now be attempted.

For Afrikaans the effectiveness of the three commercial spellcheckers, viz. Corel WordPerfect, Pharos, and Potchefstroom, was tested on a variety of randomly selected texts. For the purpose of this paper only a brief summary of the outcome will be offered, exemplified on a small section from the *White Pages*, as shown in Table 1.

Table 1: Spellchecking a randomly selected Afrikaans section from the *White Pages* (1999-2000: 14)

Afrikaans spellchecker	**Number of words in sample**	**Number of correct words *not* recognised**	**Success rate**
Corel WordPerfect	203	11	94.6%
Pharos	203	7	96.6%
Potchefstroom	203	11	94.6%

From Table 1 it is clear that the overall percentage of error detection is quite acceptable. From the subsequent experiments, however, it became clear that all three spellcheckers do not fare well with the numerous *compounds* characteristic of the Afrikaans language, a problematic situation from a users' point of view. This thus reflects the 'limits' of first-generation spellcheckers for Afrikaans based on top-frequency wordlists.

Turning to the African languages, tests were conducted on two randomly selected paragraphs, (1) and (2) below. A single glance at these texts immediately reveals that isiZulu has a conjunctive orthography while Sepedi is written disjunctively. In (1) the isiZulu paragraph is shown, where the word-forms in bold are not recognised by the Corel WordPerfect spellchecker software.

(1) Spellchecking a randomly selected Zulu paragraph from *Bona Zulu* (June 2000: 114)

Izingane **ezizichamelayo** zivame ukuhlala **ngokuhlukumezeka** kanti akufanele **ziphathwe** ngaleyondlela. Uma ushaya ingane ngoba **izichamelile** usuke **uyihlukumeza** ngoba lokho **ayikwenzi** ngamabomu njengoba iningi labazali **licabanga** kanjalo. Uma nawe **mzali usubuyisa** ingqondo, usho ukuthi ikhona ingane **engajatshuliswa wukuvuka** embhedeni obandayo **omanzi** njalo ekuseni?

The stored isiZulu list consists of the 33,526 most frequently used word-forms. As 12 out of 41 word-forms were not recognised in (1), this implies a success rate of 'only' 70.7%.

When we test the Corel WordPerfect spellchecker software on a randomly selected Sepedi paragraph, however, the results are as shown in (2).

(2) Spellchecking a randomly selected Sepedi section from the *White Pages* (1999-2000: 24)

Dikarata tša mogala di a hwetšagala ka go fapafapana goba R15, R20, (R2 ke mahala) R50, R100 goba R200. Gomme di ka šomišwa go megala ya **Telkom** ka moka (ye metala) Ge tšhelete ka moka e fedile **karateng** o ka tsentša karata ye nngwe ntle le go šitiša poledišano ya gago mogaleng.

Even though the stored Sepedi list is smaller than the isiZulu one, as it only consists of the 27,020 most frequently used word-forms, with 2 unrecognised words out of 46, the success rate is as high as 95.7%.

In an extensive second series of experiments the aim was to establish the error-detection power resulting from the cumulative build-up of top-frequency wordlists as the basis for spellcheckers for these languages. It was found that Sepedi reaches an acceptable success rate with a much smaller word list than for Afrikaans and that the success rate for isiZulu is lower even when very large word lists are used.

From a users' perspective, the success rate of a first-generation spellchecker for a conjunctively-written language like isiZulu is not really acceptable. Disjunctivism is however a great advantage for isolated-word spellchecking, as is clear from the Sepedi data. For Afrikaans, large wordlists can 'just' do.

5. Towards second-generation spellcheckers

From the above it follows that advanced technologies, as for English, for example, should be developed in what we prefer to call second-generation spellcheckers for Afrikaans, to cater for compounds in another way than the mere stacking of words in a wordlist. First-generation spellcheckers for Afrikaans could thus be improved by programming sets of rules for compounding. A spellchecker based on a true morphological analyser / generator of the language is, however, the ideal solution.

For the African languages it is clear that, for isolated-word spellchecking purposes of the Nguni languages, second-generation spellcheckers are needed to reach a more satisfactory rate of error detection. With this in mind, a thorough study was undertaken of the degree of conjunctivism / disjunctivism of all official South African languages. The results of this endeavour are shown in Table 2.

Table 2: Degrees of conjunctivism / disjunctivism for the South African languages (based on counts derived from 55 two-by-two parallel corpora, cf. Prinsloo & De Schryver 2002: 261)

	isiNdebele	Siswati	isiXhosa	isiZulu	English	Afrikaans	Xitsonga	Setswana	Tshivenda	Sepedi	Sesotho
isiNdebele	**1.00**	1.01	1.01	1.04	1.41	1.41	1.61	1.63	1.67	1.73	1.77
Siswati	0.99	**1.00**	1.03	1.04	1.41	1.41	1.61	1.62	1.69	1.72	1.77
isiXhosa	0.99	0.97	**1.00**	1.01	1.36	1.37	1.58	1.58	1.75	1.67	1.71
isiZulu	0.96	0.97	0.99	**1.00**	1.32	1.34	1.54	1.55	1.58	1.60	1.66
English	0.71	0.71	0.74	0.76	**1.00**	1.00	1.15	1.16	1.19	1.24	1.25
Afrikaans	0.71	0.71	0.73	0.75	1.00	**1.00**	1.15	1.16	1.19	1.23	1.24
Xitsonga	0.62	0.62	0.63	0.65	0.87	0.87	**1.00**	1.01	1.05	1.06	1.08
Setswana	0.62	0.62	0.63	0.64	0.86	0.86	0.99	**1.00**	1.03	1.07	1.08
Tshivenda	0.60	0.59	0.57	0.63	0.84	0.84	0.96	0.97	**1.00**	1.03	1.08
Sepedi	0.58	0.58	0.60	0.62	0.81	0.81	0.94	0.94	0.97	**1.00**	1.02
Sesotho	0.57	0.57	0.58	0.60	0.80	0.80	0.92	0.92	0.93	0.98	**1.00**

From Table 2 one can for instance see that Sepedi is 60% more disjunctive than isiZulu, or that isiNdebele is 57% more conjunctive than Sesotho. The figures in this table have a *direct* impact on the success rate of spellcheckers for the African languages, as a higher degree of conjunctivism implies a lower degree of success rate.

In contrast to Afrikaans, the main error-detection problem for the African languages is not one of compounding, but one of morphophonological changes resulting from the agglutination of morphemes in especially the Nguni languages. It is thus suggested that a proper morphological analyser / generator be incorporated into the second-generation spellcheckers for the African languages – and finite-state tools are indeed already being developed to this end, cf. e.g. Bosch & Pretorius (2002) for isiZulu, or De Schryver (2002b) for Sepedi.

Looking ahead, to the third-generation spellcheckers, these will of course need to have a grammar component as well. For the disjunctively-written languages this will for instance 'solve' the current problem that a correct sequence of two or more equal words is marked as potentially wrong.

6. Conclusion

We have seen that first-generation spellcheckers, viz. spellcheckers based on top-frequency wordlists, result in *just acceptable* error-detection software for a language like Afrikaans which is characterised by extensive compounding. Conversely, this same approach produces *excellent* error-detection software for disjunctively-written South African languages. For the conjunctively-written South African languages (the Nguni languages), however, even long lists of word-forms can *not really* be considered *acceptable*. The success of isolated-word error detection for the African languages is inversely related to the degree of conjunctivism.

It was further suggested that the second generation of spellcheckers for Afrikaans include some basic compounding rules, and that the conjunctively-written South African languages include a morphological analyser / generator. The latter will of course be a crucial component of *all* South African third-generation spellcheckers – spellcheckers which will also be able to perform grammatical checks.

References

Bona Zulu, Imagazini Yesizwe, Durban, June 2000.

Bosch, S.E. and L. Pretorius. 2002. Using Finite-State Computational Morphology to Enhance a Machine-Readable Lexicon. In G.-M. de Schryver (ed.). 2002a: 20-22.

De Schryver, G.-M. (ed.). 2002a. *AFRILEX 2002, Culture and Dictionaries, Programme and Abstracts*. Pretoria: $(SF)^2$ Press.

De Schryver, G.-M. 2002b. First Steps in the Finite-State Morphological Analysis of Northern Sotho. In G.-M. de Schryver (ed.). 2002a: 22-23.

Hurskainen, A. 1999. SALAMA: Swahili Language Manager. *Nordic Journal of African Studies* 8/2: 139-157.

Kukich, K. 1992. Techniques for Automatically Correcting Words in Text. *ACM Computing Surveys* 24/4: 377-439.

Prinsloo, D.J. and G.-M. de Schryver. 2001. Corpus applications for the African languages, with special reference to research, teaching, learning and software. *Southern African Linguistics and Applied Language Studies* 19/1-2: 111-131.

Prinsloo, D.J. and G.-M. de Schryver. 2002. Towards an 11 x 11 Array for the Degree of Conjunctivism / Disjunctivism of the South African Languages. *Nordic Journal of African Studies* 11/2: 249-265.

White Pages Pretoria, North Sotho – English – Afrikaans Information Pages, Johannesburg, 1999-2000.

Web Services in Language Technology and Terminology Management

Uwe QUASTHOFF & Christian WOLFF
Computer Science Institute, NLP Department, Leipzig University, Germany

Abstract: In this paper we describe the application of Web services towards language technology and terminology management. Starting from a short review of Web development, the notion of Web services is introduced and relevant standards in this area are briefly described. Following a motivation for converting language technology applications into Web services we give examples of such services based on a large language and terminology service developed over the last few years. The question of modelling language technology services is discussed as well. Finally, some technical details illustrate our Web service prototype.

1. Introduction

If we compare a terminological database with written resources, we find the following well-known advantages:

- Size: Usually a database contains much more data than the books in a typical bookshelf.
- Quick look-up: Looking up in a database is quicker than finding an entry, even in one dictionary. This becomes even worse if one has to look up in several dictionaries.

During the last few years several useful databases were created. While many areas of language technology and terminology management have been covered so far, some shortcomings of this approach have become obvious as well:

- Similar steps have to be repeated to look up information on the same words or concepts in several databases.
- Databases from different vendors have different user interfaces.
- Different databases may have different data structures or different query capabilities.

In this paper, we describe a Web service-based approach for overcoming these disadvantages. Web services can be used for a standardised and unified access to terminology information using a single user interface. The terminology vendor only provides the data via the Web service. Using an automatic mapping of the database structure, the presentation in the user interface can be customised to the user's needs. Moreover, several databases with different internal structures can be presented in a single way.

We illustrate the methods using the Web service of the various databases at the *Leipzig Wortschatz* at http://wortschatz.uni-leipzig.de (see Quasthoff & Wolff (2000) and Heyer et al. (2002) for more information on this project).

2. The Web: From static hypertext to modular Web services

2.1. Web development

In its short, 10+ year, history the Web has seen dramatic technological change. It may roughly be categorised into three major phases of development (see also Preece & Decker (2002: 15) who propose a simpler, two-phase development model):

1. Initially the Web was designed as a means for distributing hypertext based on a simple protocol (HTTP). The Web browser was introduced as client software.
2. The second phase, starting during the later part of the nineties, brought information systems to the Web. Web information systems allow for the presentation of arbitrary information system functionality via common standards and common client software (the ubiquitous Web browser).
3. The third phase, which has begun only very recently, introduces two additional innovations. On the one hand, enriched languages are used on the Web by providing meta-information standards like the Resource Description Framework (RDF) or the Topic Map ISO-standard. The aim is a vision of the "semantic Web" which makes more complex Web-based applications possible by better resource description. The second development, which is described in more detail in this paper, is the generalisation of Web-based functionality as Web services. While Web information systems employ a broad variety of not necessarily compatible technologies, Web services make functionality and information available via standardised descriptions and access protocols.

All three stages of Web development have been picked up by the language technology and terminology management community:

1. Terminology lists and dictionaries have been published as static hypertext on the Web for many years (phase 1).
2. More recently, language technology applications like stemming, machine translation or information extraction have become available as Web information systems (phase 2).
3. Finally, the language technology and terminology management community is beginning to transform Web information systems into more flexible Web services (phase 3).

A good overview of available resources may be found at *Language Technology World* (http://www.lt-world.org/ns_index.html), a comprehensive repository for all kinds of language technology-related resources and applications.

2.2. Standards for Web services

A key issue in Web service development is standardisation, as only a common ground for Web services will allow essential features of the third phase of Web development to be achieved:

- Modular composition of different Web services.
- Integration of Web services into complex applications.
- Universal access to Web services from different types of client software and client applications.

Currently, three standards have been proposed which address key aspects of Web services and which are also employed in the language technology examples given below:

- SOAP, being a simple messaging protocol for actual Web service deployment (cf. Gudgin et al. 2002).
- UDDI (Universal Description, Discovery and Integration), being a standard for Web service directories and Web service lookup (cf. Bellwood et al. 2002).
- WSDL (Web Service Description Language), providing a common framework for the abstract description of Web service functionality (cf. Chinnici et al. 2002).

From a technical perspective, Web services reinvent the traditional idea of remote procedure calls for the Web, as arbitrary functionality can be accessed via the Web. From a (naïve) users' perspective, a Web service simply returns information to a given question.

For the client side, a so-called generic SOAP client is available which allows instant access to an existing Web service via a browser. Especially in the case of language services, the service provider can maintain a large database locally and the user has always access to the current data. In a user-driven approach, we can specify the data wanted in a typical dictionary lookup. This will give us a set of useful methods one wants to have.

From the providers' point of view, these methods have to be implemented for their database. This hides the actual database structure and all the related technical details from the user. Additional practical examples of existing language-related Web services may be found at http://www.remotemethods.com/home/valueman/converte/humanlan.

3. Web services for language technology and terminology management

In this section, we want to describe methods which are useful for terminology lookup and terminology generation. Examples are taken from the *Leipzig Wortschatz* project mentioned above which comprises:

- a large text corpus;

- a comprehensive dictionary of inflected forms with a rich data structure for each entry (statistical information, semantic attributes, morphological and syntactical information);
- additional features extracted from text via text mining like collocations for each entry;
- a rich set of tools for corpus and dictionary set-up, analysis, and maintenance.

3.1. Query types

Different Web service methods may be categorised either structurally with respect to the underlying database model developed, or according to the information need modelled by a Web service method. As the structural aspect is an inherent technical one, we will concentrate on different typical information needs only.

3.2. Full dictionary lookup: `give_entry`

The `give_entry` method returns the "classic" dictionary entry for a given term. The whole entry is returned as a block of text in XML format with XML tags delimiting (and describing) the logical parts of the dictionary entry. While this is useful for typical dictionary (or terminology database) usage by terminologists, the flexibility of a composable Web service is not fully exploited. There are several reasons one may wish to get only parts of a dictionary entry, using more specific or atomic rather than composite Web service methods, as will be shown in the next section.

3.3. Partial dictionary lookup

Linguistic databases can contain a lot of information about a single word: The monolingual part may contain statistical, grammatical and semantic information. There may be additional multilingual parts.

In the translation process one might be interested in a specific language pair and, moreover, subject area information contained in the monolingual part. Only these fields are relevant. Hence, we are able to define special Web service methods which give just the desired fields. A very simple example of such an atomic Web service method is given in the Appendix where a SOAP example of a `getBaseForms` Web service method for the *Leipzig Wortschatz* database is shown.

3.4. Terminology extraction

In addition to dictionary lookup, text analysis is an interesting application for Web services. Results can be monolingual terminology lists derived from the text given. Combined with bilingual resources, also a bilingual terminology list can be produced. An example of this kind of Web service is the *Concept Extractor*, a Web service-based software tool developed on top of the *Leipzig Wortschatz* infrastructure which extracts

relevant terminology from given texts via an application of differential corpus analysis (see Faulstich et al. (2002) for further details).

4. Conclusion

Over the past few months we have started developing and offering Web services for terminological information which may be used for information presentation as well as integration into language technology applications. While, on a technological level, standards for offering such services have become available, further standardisation is needed. We are working on a complete set of Web service functions, atomic as well as composite, for the most pressing needs of our users in the language technology and terminology management area.

References

Bellwood, T. et al. 2002. *UDDI Version 3.0. Universal Description, Discovery and Integration (UDDI) Project, Published Specification, July 2002*. <http://uddi.org/pubs/uddi_v3.htm>

Chinnici, R. et al. 2002. *Web Services Description Language (WSDL) Version 1.2. World Wide Web Consortium Working Draft, July 2002*. <http://www.w3.org/TR/wsdl12>

Faulstich, L. C., U. Quasthoff, F. Schmidt and C. Wolff. 2002. Concept Extractor – Ein flexibler und domänenspezifischer Web Service zur Beschlagwortung von Texten. In R. Hammwöhner, C. Wolff and C. Womser-Haccker. *Information und Mobilität, Proc. 8. International Symposion in Information Science, Regensburg, October 2002*.

Gudgin, M. et al. 2002. *SOAP Version 1.2. Part 1: Messaging Framework. World Wide Web Consortium Working Draft, June 2002*. <http://www.w3.org/TR/soap12-part1>

Heyer, G., U. Quasthoff and C. Wolff. 2002. Knowledge Extraction from Text: Using Filters on Collocation Sets. In *Proceedings of LREC 2002. Third International Conference on Language Resources and Evaluation, Vol. III*: 241-246. Las Palmas, May 2002.

Preece, A. and M. Decker. 2002. Intelligent Web Services. *IEEE Intelligent Systems* 17/1: 15-17.

Quasthoff, U. and C. Wolff. 2000. An Infrastructure for Corpus-Based Monolingual Dictionaries. In *Proceedings of LREC 2000. Second International Conference on Language Resources and Evaluation, Vol. I*: 241-246. Athens, May / June 2000.

Appendix

SOAP Request-Response Example
SOAP source code for `getBaseForms`

```
Request:
<?xml version="1.0" encoding="UTF-8"?>
<soapenv:Envelope

soapenv:encodingStyle="http://schemas.xmlsoap.org/soap/encoding/"
  xmlns:soapenv="http://schemas.xmlsoap.org/soap/envelope/"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns:SOAP-ENC="http://schemas.xmlsoap.org/soap/encoding/">
 <soapenv:Body>
  <ns1:getBaseForms xmlns:ns1="urn:LdbApi">
   <word xsi:type="xsd:string">Sachsen</word>
  </ns1:getBaseForms>
 </soapenv:Body>
</soapenv:Envelope>


Response:
<?xml version="1.0" encoding="UTF-8"?>
<soapenv:Envelope
  xmlns:soapenv="http://schemas.xmlsoap.org/soap/envelope/"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
 <soapenv:Body>
  <ns1:getBaseFormsResponse

soapenv:encodingStyle="http://schemas.xmlsoap.org/soap/encoding/"
     xmlns:ns1="urn:LdbApi">
   <getBaseFormsReturn xsi:type="soapenc:Array"
       soapenc:arrayType="xsd:any[3]"
       xmlns:soapenc="http://schemas.xmlsoap.org/soap/encoding/">
    <item xsi:type="xsd:string">Sachsen</item>
    <item xsi:type="xsd:string">Sachs</item>
    <item xsi:type="xsd:string">Sachse</item>
   </getBaseFormsReturn>
  </ns1:getBaseFormsResponse>
 </soapenv:Body>
</soapenv:Envelope>
```

Generic SOAP Client Screenshots for the `getBaseForms` Web Service

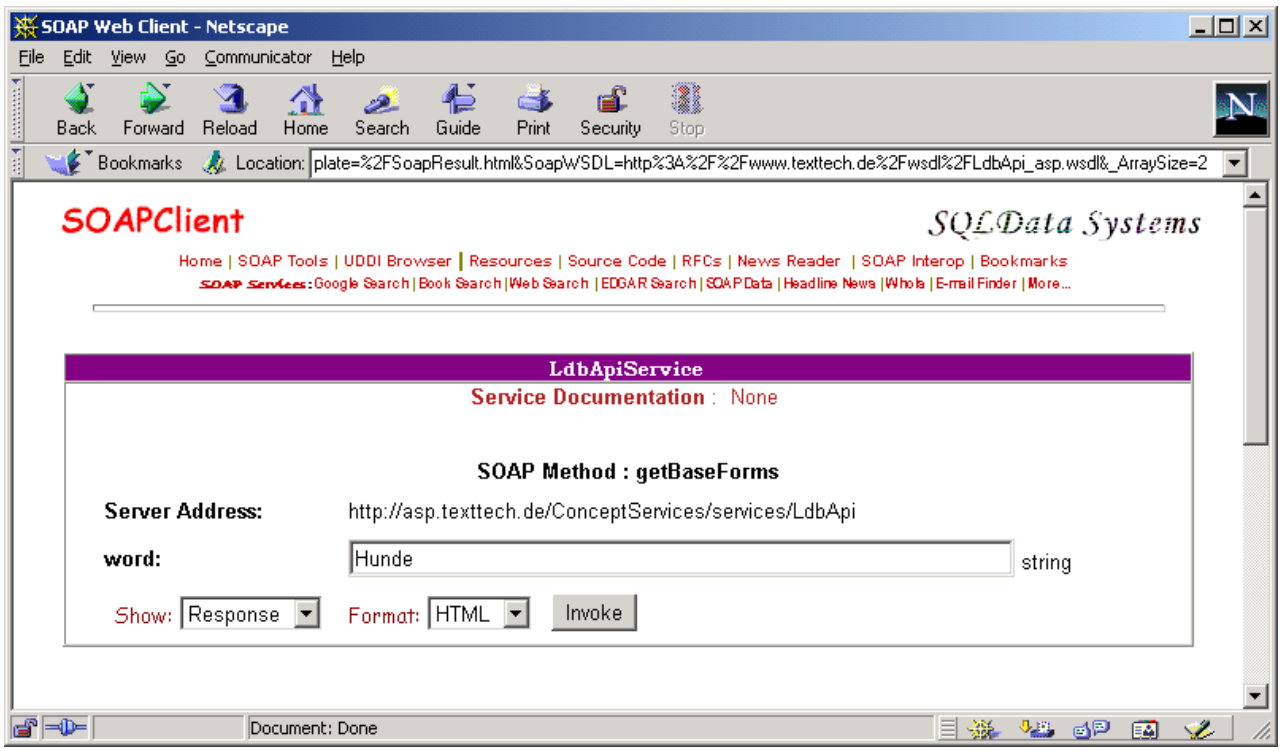

Figure 1: Generic SOAP client interface for the `getBaseForms` service

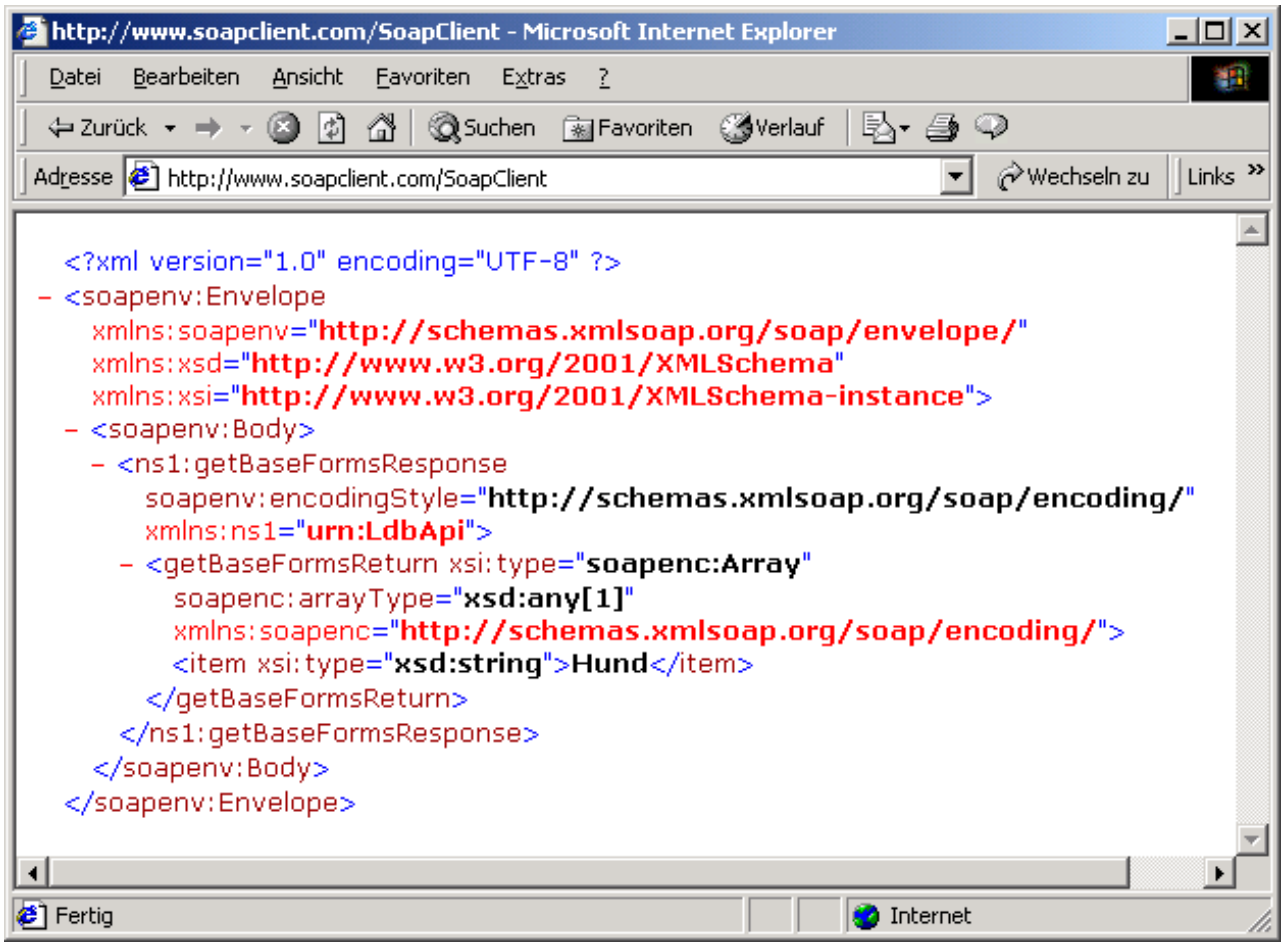

Figure 2: XML output for the `getBaseForms` service

Developing Human Language Technologies in South Africa: Challenges and Proposals

Justus ROUX
Research Unit for Experimental Phonology, University of Stellenbosch, SA

1. Aim

The aim of this paper is to reflect on some developments that have been taking place with respect to the development of *Human Language Technologies* (HLTs) in South Africa over the last four years and to focus on challenges that need to be addressed as the country moves into a technology-driven Information Society.

2. Challenges

Human Language Technologies are *enabling technologies* that:

- facilitate a process in which humans can interact with machines (computers) in the most natural manner, i.e. through normal language (text or speech), and
- render support to human operators in performing specific tasks, e.g. assisting in translating documents from one language to the other, or
- allow tasks to be performed through verbal interaction, e.g. security systems reacting to human language or speech.

The Information Society is a knowledge society in which access to information is of the utmost importance. This information is typically made available in electronic format, i.e. either through telephone connections linked to computer systems, or in text format on the Internet.

One of the *main challenges* of electronic service delivery is that of facilitating *universal access*. Universal Access refers to the right of access to information regardless of disability, economic situation, and geographic location or language proficiency.

A *second challenge* is to facilitate *intercultural communication and understanding* through the use of appropriate HLTs. This challenge presents itself at:

- formal governmental level;
- the level of business and industry;
- the level of mass communication.

A *third challenge* for HLTs relates to *education and training* in the widest context. It has become commonplace in many schools and tertiary institutions to employ e-learning strategies in computer centres. The development of interactive multimedia

learning materials across the curriculum can be enhanced by multilingual components based on HLTs.

This is by no means an exhaustive list of challenges and it is possible to enumerate various other challenges that call for the implementation of HLTs in one way or the other. It is important to note that it is generally accepted that HLTs have an extremely important role to play in bridging the *digital divide*.

3. Global perspective

Multilingual Europe has recognised the need to develop HLTs at national levels. This is to allow for people of different nationalities and languages to interact with one another through a technological medium. The EU assigned an amount of *3,295 million Euros* for ICT (which includes the development of HLT) in the Fifth Framework Programme – this underscores the importance attached to this field.

4. Activities within South Africa

It is clear from public statements and from several (uncoordinated) projects that the South African National Government is indeed aware of developments taking place in the field of HLT and of the benefits it may hold for the nation. There seems to be a commitment to employ technology to enhance *multilingualism* as well as to set up a 'language industry'.

This awareness is reflected in various public statements by, or on behalf of, National Government:

- The **DACST Foresight Project** (1999) report entitled *Dawn of the African Century*, Section 4.8 "Information and Communication Technologies", states quite clearly: *'High priority should be given to a significant level of research and technology investment in Human Language Technologies'* (p. 40).
- The **Final Draft Language Policy and Plan for South Africa** (2000) proposes the following:
  - *'Building capacity in the field of language and technology for all South African languages'* as a strategic goal (p. 8).
  - *'Development of an efficient language industry by, among other things, using and developing appropriate technology'* (p. 9) and *'Supporting language technology'* (p. 10).
  - *'Government shall encourage and, wherever necessary, support the development of language technology for South African languages'* (p. 16).
- November 1999: PanSALB endorsed that the Translation and Interpreting Subcommittee (TISC) should pursue the Human Language Technologies (HLT) Project. This resulted in the establishment of a **Joint Steering Committee (PanSALB & DACST) on HLT**, with the task to draft a Strategic Plan for the development of HLT in South Africa. This report was submitted in 2000.

- In response to this report the Minister of Arts, Culture, Science and Technology appointed an **Advisory Panel on HLT** with the task to determine practical implications and cost factor of such an initiative. The Panel submitted its report in September 2002.
- The necessity for the development of HLTs has also been emphasized in a policy framework of the *Department of Public Service and Administration* (DPSA) entitled **Electronic Government: The Digital Future** (2001), inter alia calling on DACST to take the lead in implementation.
- The first steps have already been taken to implement an e-government strategy in South Africa through the **e-Government Gateway Project** of DPSA, in which the implementation of HLTs is of core importance to ensure adequate access to information in a multilingual society.
- The new **National Research and Development Strategy of SA** (2002) makes explicit reference to the importance of the development of Human Language Technologies within the ICT domain.
- The importance of the development of HLTs was emphasised by the Director General of ACST at the Human Language Technology Launch, in 2001: *'... an innovative attempt at building capacity in HLT and implementing technology to promote multilingualism and to fast track the development of the marginalised indigenous languages.'*

Over and above the intent of National Government to develop the field, a number of HLT-related projects are currently being conducted:

- The DACST Innovation Fund (Round 2) is funding a major project entitled **African Speech Technology**. This Stellenbosch University project is conducted by a local team with more than eighty co-workers from across SA of which more than half are mother-tongue speakers of an African language.
- The **Multilingualism, Informatics and Development Project** is a partnership project between the province of Flanders, the Universities of Antwerp and the Free State, as well as the Provincial government of the Free State.
- The **Telephone Interpreting Service for South Africa** (TISSA) of the *Department of Arts & Culture* is a project of the *National Language Service*, the CSIR and the Unit for Language Facilitation and Empowerment (Free State University).
- The **Development of Spellcheckers for the Indigenous Languages of South Africa** by Afrilex, within the Department of African Languages at the University of Pretoria.
- The **Development of Morphological Parsers and Tools for the African Languages** comprises three projects conducted by UNISA with the cooperation of

the *Special Interest Group for Speech and Language Technology Development* of the *African Language Association of SA*.

4.1. Examples of typical potential HLT-driven systems in the South African context

- Multilingual telephone based information systems:
  - Tourism & Travel: Hotel booking systems (AST project); train, air, bus schedules; road conditions, weather reports, travel packages.
  - Public services: Applications for pensions, travel documents, car registrations; telephone accounts, telephone number enquiries.
- Multilingual multimedia information systems:
  - Education: Language learning, voice-based training systems.
  - For non-literates: Using voice to obtain information automatically.
- Multilingual automatic translation systems / Translation Memory systems:
  - State services: Official documents, Hansard in national, provincial, local governments.
  - Education: Developing multilingual teaching material.

5. Addressing the challenges

Addressing the challenges could be accomplished by responding to the findings of the PanSALB & DACST Steering Committee Report (2000). This report revealed a need for:

- Reusable language (text and speech) resources in electronic format.
- Capacity building in the field of Human Language Technologies, including Lexicography.

Proposals that were made in the Steering Committee Report (2000) included:

- The establishment of a National Centre for Electronic Text and Speech.
- A network linking the *National Lexicography Units* (NLUs) to the Centre.

5.1. Why a National Resource Centre?

- Need to co-ordinate and systematise the massive amounts of multilingual electronic text and speech data generated daily in SA from various sources.
- Need to develop appropriate software to access this data from these corpora and to convert it into a format usable to potential users, i.e. HLT developers as well as lexicographers.
- Need for systematic sustainable training programmes, inter alia in lexicography and terminology.

The Ministerial Advisory Panel (2002) refined the ideas expressed in the 2000 Report. Although this report is not yet in the public domain, and Government has not yet made

any decision, the idea of establishing a **National Centre for Human Language Technologies** was widely discussed in public and created keen interest among different major role players including the *National Research Foundation* (NRF) and the SABC.

5.2. The Nature of a National Centre for HLT

The proposed centre will be *virtual* in nature and the following actions / tasks are foreseen:

- The identification of a physical "hub" in South Africa at a centre where some HLT expertise is available in the following disciplines: (African) linguistics, electronic engineering and computer science. This hub will be linked electronically to various nodes at other universities, technicons, etc. which will participate in accordance with set guidelines and priorities.
- A joint national panel of relevant role players and an International Advisory Panel of HLT experts will determine the activities and priorities of the Centre.
- The centre will have *three basic activities*:
  - Act as a *depository for applicable electronic text and speech data* that will be developed, managed and distributed according to a proven international model. The *reusable resources* that will be created will include Sign language materials.
  - Develop applicable *Natural Language Processing (NLP) software*, specifically to address the needs of the African languages. This *open sourced software* will assist in fast tracking of the technological development of these languages.
  - Identify *human capacity needs* with respect to HLT skills and devise custom-made *training and/or reskilling programmes* to meet these needs.

5.3. Organisational structure of a National Centre for HLT

It is not yet possible to determine the most suitable organisational and/or staff structure, however, this centre should:

- strive to ensure maximum participation of all role players as producers and/or users of relevant digital resources;
- endeavour to keep the management component as small, but as effective, as possible;
- promote growth in research at tertiary institutions as well as the establishment of SMEs with the cooperation of local and national incubation centres, e.g. with the CSIR.

5.4. Physical infrastructure of a National Centre for HLT

This will be a *virtual centre* with the idea to negotiate partnerships with academia and/or business and industry (inter alia, for continued funding). The underlying

philosophy is that the work should be done by experts already working in the field, and that their infrastructures may be used for further research and training (with student support from the NRF). In order to prevent duplication in the development and distribution of costly digital data it is absolutely necessary that central planning and co-ordination should be the corner stone of this action. The diagram shown in Figure 1 demonstrates the various relationships:

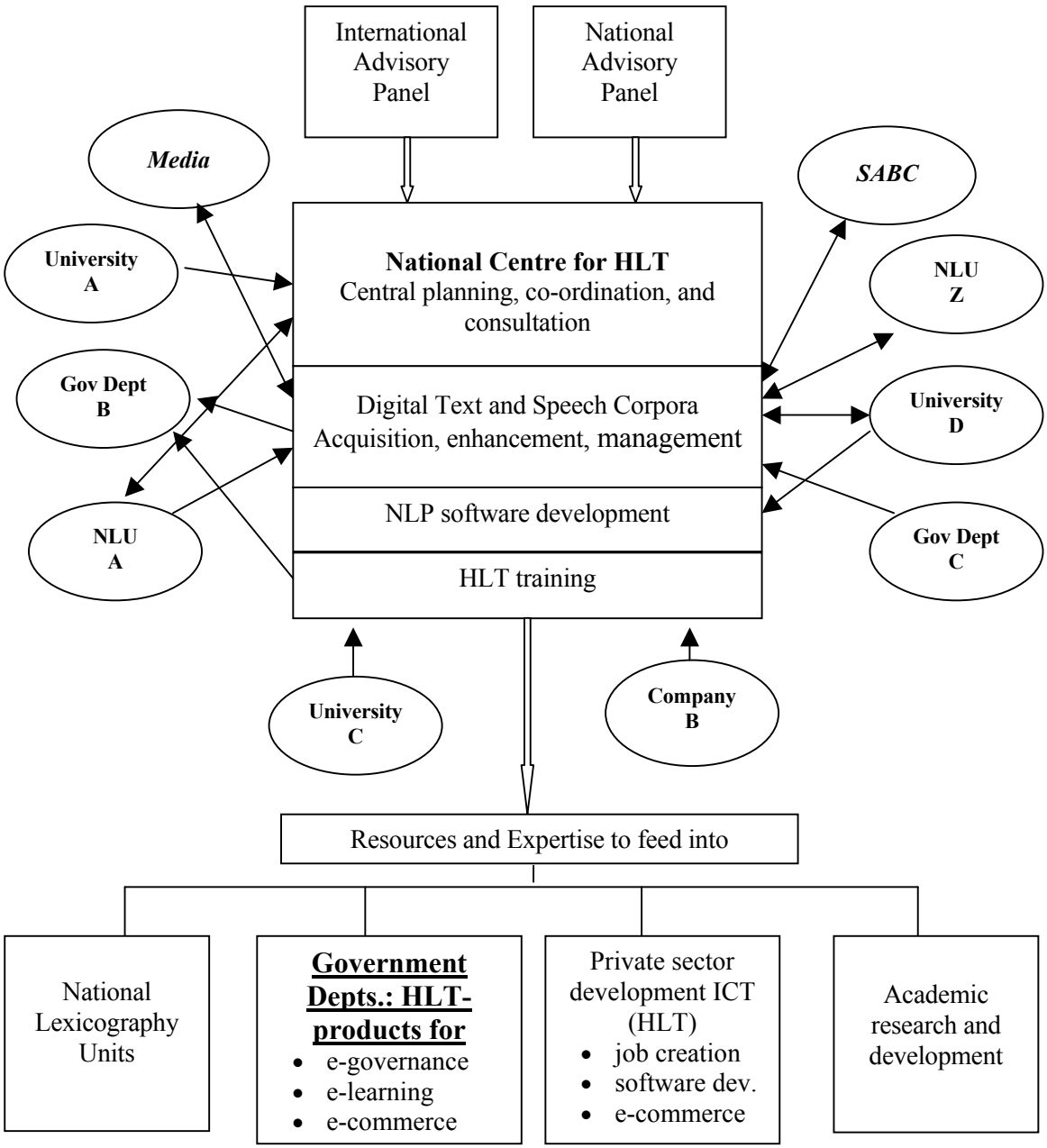

Figure 1: Diagram of the proposed *National Centre for Human Language Technologies*

6. Concluding remarks

I am of the opinion that South Africa is in a unique position to develop its HLT capacity. Whilst we may be in a position to learn from the (technological) mistakes made by others in the past, we should also join this field as soon as possible with innovative projects and strategies. We are potentially on the verge of a new approach to lexicography, terminology and language and speech technologies. It is, however, imperative that we see the big picture and integrate all components conceptually. It is equally important to note that the development of terminology has become part and parcel of this field. The importance of terminology within the field of Human Language Technologies is clearly demonstrated by the growing number of contributions to a series of dedicated International Conferences on *Language Resources and Evaluation* (LREC), which respectively took place in Granada (1998), Athens (2000) and Las Palmas (2002).

References

LREC. 1998. *Proceedings of the First International Conference on Language Resources and Evaluation, Vol. I & II*. ELRA: Granada.

LREC. 2000. *Proceedings of the Second International Conference on Language Resources and Evaluation, Vol. I, II & III*. ELRA: Athens.

LREC. 2002. *Proceedings of the Third International Conference on Language Resources and Evaluation, Vol. I, II, III, IV, V & VI*. ELRA: Las Palmas.

Report. 2000. *The development of HLT in South Africa*. <http://www.dacst.gov.za/arts_culture/index.htm>

Asynchronous Learning Environments for Language, Communication and Culture Study

Irina N. ROZINA[°], Ronald D. ECKARD[‡] & Joe DOWNING[#]
Department of Information Technologies, Institute of Management, Business and Law, Rostov-na-Donu, Russia[°], Department of English, Western Kentucky University, USA[‡] & Department of Corporate Communications, Southern Methodist University, USA[#]

Abstract: Intercultural dialogue in contemporary global society is proceeding with amazing speed as a result of the advancement of synchronous and asynchronous telecommunication technologies. Electronic dialogue users need to be prepared to use such technologies as e-mail, WWW-based technologies, forums, IRC, etc. Scholars are challenged to develop multilingual knowledge and new terminology training, to improve innovative teaching and learning methods so as to properly and efficiently hold intercultural dialogues.

Here we present some successful Correspondence, Cultural Values and Communication projects for language, communication and culture run at some Russian and US Universities from 2000 through 2002, based on asynchronous technologies (basically e-mail, and Web-form for the Cultural Values survey). The results of these projects confirm our pedagogical hypothesis that students from different countries can be themselves language, communication and culture teachers, as well as a source of appropriate knowledge in asynchronous learning environments based on e-mail and Web technologies.

1. Introduction

Cross-cultural interaction in modern global society is proceeding with amazing speed as a result of advancements in asynchronous and synchronous telecommunication technologies. This means that educational organisations are challenged to teach their students how to use new computer-mediated communication technologies (e-mail, WWW, forums, IRC, etc.) and to research how to develop them properly and efficiently in education. Here we present some successful methodology, which was introduced in joint Correspondence, Cultural Values and Communication projects aimed at language, communication and culture learning.

The project was carried out by the Institute of Linguistics, Rostov State Pedagogical University, Russia; the Philology Department of the Institute of Management, Business and Law, Russia; and the English and Communication Departments of Western Kentucky University, USA, from 2000 through 2002. The educational activity, which we called an educational project, helped the students to develop their skills in using telecommunication technologies, and connected them and gave them a more appropriate medium for language and culture learning.

2. The theoretical aspect

The main hypothesis underlying these projects comes from our understanding that students from different countries can themselves be language, communication and culture teachers and a source of appropriate knowledge. This hypothesis is built on the best known of Lev Vygotsky's social-theoretical concepts of proximal development zone (ZPD). According to Vygotsky the culture gives students the cognitive tools needed for their development. Foreign students are conduits for the tools of the foreign culture, including language, terminology and communication differences. The tools the culture provides, include social context, language and asynchronous telecommunication technologies for interactive communication and information exchange. In turn, contemporary telecommunication technologies bring additional methods, tools, common terminology and criteria for successful learning environment development. Thus we gradually see that the evolution of a learning environment is based both upon theoretical and practical elements.

The first thing that helps educators in this new learning environment is that learners become more active participants. The second is that the learners accept and incorporate previous and current learning practices in terminology from different fields such as language, computing and telecomputing that rely upon emotional and social factors. Thirdly, many successful learning tools were initially not aimed at learning: word processors, spreadsheets, simulations programs, expert systems, e-mail, mailing lists and all WWW-based communication tools. In our project virtual learning communities were developed with the goal to activate these tools for new learning environments.

Effective learning is a many-faceted process that benefits from these new tools. Generally, positive characteristics of these application tools for effective learning are as follows:

- They stimulate active intellectual involvement on the part of the learner.
- The student, rather than the teacher, is in charge of the learning activity – students determine when and how they use computing and telecomputing to support their efforts to present information, collect data, solve problems, collaborate, and persuade others or foreign respondents.
- The learners have control of communication via e-mail. Students prefer to use a word processor with grammar and spelling checkers, the thesaurus, vocabularies, various editing aids such as replacements, block moves, copying, citing, etc.
- The correspondence project methods are aimed at accomplishing more creative tasks than traditional methods of teaching language, communication, culture, computing and telecomputing methods.
- These learning media assimilate new genres, styles of writing and terminology (handling slang, emoticons, abbreviations, signatures, etc), and new network etiquette (netiquette) in order to express meanings.

- Collaboration using learning media promotes mutual tolerance and provides responsibility to make students from both countries conscious of their identity and their responsibility to help one another to learn culture, communication and language as well.

Transcripts were collected from a graduate level virtual conference seminar course taught during a 1998 summer session at Northern Arizona University. Participants were in-service classroom teachers. A kindergarten through high school grade range of classroom assignments was represented by these teachers. Course work included outside class reading assignments of selected current research in the field, a final research paper, and active participation in the online discussion forum with focus questions by the instructor. The seminar offered was a Tools for Teachers course designed to promote reflective practice. Seven teachers from four different communities participated in this pilot course during the summer of 1998. Five of the participants were female and two were males. An interactional sociolinguistics approach was used to examine the texts of conversations. This approach draws upon concepts of culture, society, language, and the self. The meaning, structure, and use of language are socially and culturally relative. Meaning in dialogue, like that of conversation, is socially constructed. Data was also examined for evidence of micro-displays that are commonly associated specifically with either gender. An example of this is the use of tag questions. Female participants, much more so than males, tend to use tag questions as a discourse strategy to invite response and inclusion or solidarity within the group.

3. The practical approach

The role of teachers from all classes was to help learners with their communication, culture and language skills. Learners periodically sent their teacher and/or partner teacher a request or a copy of their e-mail to solve communication problems, or simply to show their progress and to receive help. The pedagogic (cultural and educational) aim is:

- to encourage learners to exchange cultural information;
- to comment / "correct" the language of each other as a form of mutual self-help among students (peer collaboration).

Before the beginning of the Correspondence Project e-mail partners were given some guidelines from their virtual class teachers. For example, the guidelines adapted from the *International E-mail Tandem Network* (http://www.slf.ruhr-uni-bochum.de/e-mail/help/helpeng01.html#questions). The same approach was used in the Correspondence Project to create learning media between students from Rostov State Pedagogical University and Valenciennes University, France for Russian and French language study.

Brainstorming was one of the methods we used to prepare ideas and materials for the Correspondence Project. For example, we encouraged our students to brainstorm about their impressions and knowledge of the other country. Then we provided our students with the results of the Internet survey about the traditional impressions, or the list of perceptions, prepared in advance at partner class or by means of our Cultural Values survey. We asked them to compare results and talk about some topics.

Thus in a class at Rostov State Pedagogical University, the students were asked to brainstorm and express their perceptions of Americans and to freely express what they know about the US. Some were repetitions and some were contradictions. We discussed the students' own lists in order to record and categorise their impressions. If we had enough answers from a partner class, the most interesting pen pals' responses were shared aloud in class, then discussed. Some cultural information as to why these differences might exist was given. We then divided their responses into positive and negative descriptions. In our current project we listed all terms in order of frequency; therefore, in the last version of the database those terms listed first or second were provided by more than one student.

In the Cultural Values Project both the Russian and the American students filled in the surveys (hard, soft or Web-based form: http://rspu.edu.ru/li/cultural_value/survey_e.html) and exchanged their findings (personally forwarded to partners or collected as full database of all student responses). In the Cultural Values survey database some values were rated the highest on the Russian responses and certain common or different on the American responses. We came up with a few questions in class for them to ask their pen pals. The teacher helped them with providing some information on the types of things that might be different when comparing the Russian and American cultures. Students discussed results and asked questions based on their partner's most important values and full database results.

In the Communication Project the goal of the assignments was to analyse the mediated (e-mail) communication (computer-mediated communication, CMC) with colleagues from a different culture – American and Russian. To do this, students analysed their e-mail conversations by applying interpersonal and CMC theories we had discussed in class. Teachers paired up students so as to maintain e-mail contact between students. Usually at least three to five interchanges occurred between partners. If a partner did not respond, teachers tried to assign their student a different partner. Students saved copies of all e-mail correspondence (including copies of messages they sent out). Printouts of these e-mails were included in the students' final paper. For example, American students used any three or more of the communication theories: Norm of Reciprocity, Differences between Men and Women in Communication, Social Presence Theory, Media Richness Theory, and Hypersonal Interaction Theory. A final paper identified the theory that they used to analyse the communication with their e-mail partner, and gave a brief overview of the

theory. Students had to determine what communication behaviour the theory seeks to explain or describe, using examples taken directly from the collected e-mails to either confirm or disprove the specific tenets of each theory.

Usually students conducted each e-mail exchange for one semester. Once each student had a partner, we encouraged them to exchange e-mails as often as possible. Many of them corresponded more than once a week. We officially met in class once a week, and teachers mentioned the time and day of our meeting in a mass e-mail to both the American and Russian students, so that they might try to send their replies before the beginning of our class period each week.

We conducted the class in both a formal and informal manner. Therefore, students often felt free to disclose and then revise their initial impressions and correspondence problems. As the students received responses, they would share them with the group. The responses were usually read aloud. We are not sure if they ever revised their initial impressions, since they didn't share many of these original impressions in the lessons.

Simple quantitative analysis was done to determine computing and telecomputing terminology in the texts generated by each participant, as well as the amount of participation, total number of questions, statements, and number of responses sent and received. Patterns of participation were mapped / graphed and correlated to the contexts of interactions. Style, register, and 'voice' or tone analysis, were also used on the data sets to try to discover the dynamics among the participants and common communication terminology.

4. Conclusion

As a result of these projects, Russian students increased their English vocabulary, positively changed their perceptions about Americans and American culture, and began to practice appropriate netiquette even during the course of these one-semester-long projects. The results of this study confirm our hypothesis that students from Russia and the US can themselves be teachers of language, communication and culture, through an asynchronous learning environment design based on e-mail and Web technologies and common computing and telecomputing terminology.

DTP – the German Terminology Portal

Klaus-Dirk SCHMITZ
Institute for Information Management, University of Applied Sciences, Cologne, Germany

The German Terminology Portal (DTP = *Deutsches Terminologie-Portal*) is a Web-based information portal (http://www.iim.fh-koeln.de/dtp/) aimed at providing industry, small and medium-sized enterprises, public authorities, institutions, research centres, as well as teaching and training institutions, with various kinds of information, documentation and services that are related to terminology and the German language.

The portal mainly offers:

- Online documents and guidelines explaining the foundations of terminology theory and terminology management as well as argumentation papers.
- A selected bibliographical list with further reading related to terminology and access to a bibliographical database with terminological literature in German.
- An inventory of teaching and training opportunities as well as institutions in the field of terminology.
- A detailed inventory of terminological resources accessible via the Web and containing German terms and definitions (in addition to other languages).
- A calendar of terminology events.
- Descriptions of organisations, institutions and projects dealing with terminology in German speaking countries.
- An overview of software tools supporting terminology work.

The portal is designed and implemented during a two years project (October 2001 – September 2003), is financed by the *Ministry for Education, Science and Research of North-Rhine-Westphalia*, and is carried out by the Institute for Information Management of the University of Applied Sciences in Cologne, Germany. Partners from industry are supporting the project team with application-oriented proposals and are evaluating the design, the first prototype and the final system implementation. A close cooperation with a similar project for the Dutch language (VIPTerm) and with the Board and the members of the German Terminology Society (DTT = *Deutscher Terminologie-Tag*) is improving the results of the project.

During the presentation we will report on the project and its results, and we will also present a detailed description of the information types and the user interface of this German Terminology Portal DTP. Figures 1 to 10 below provide a first overview.

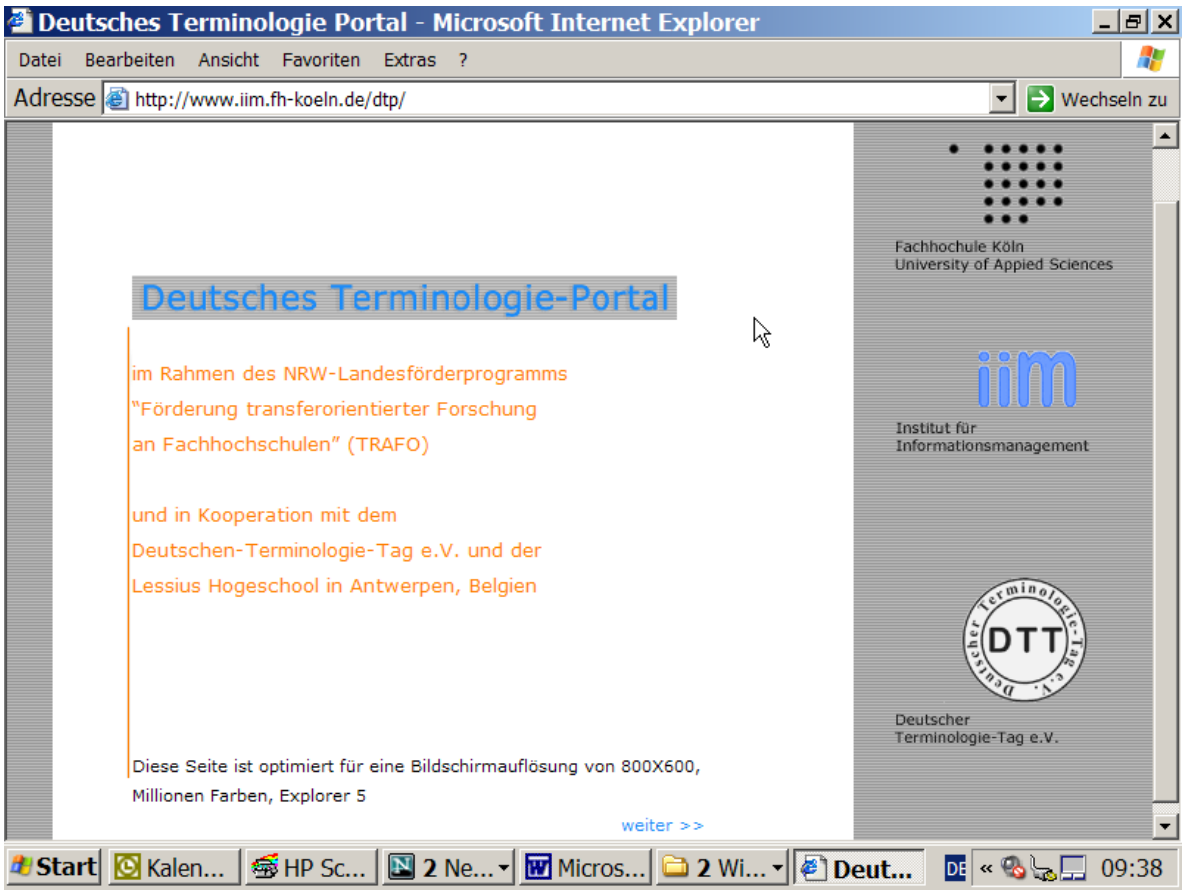

Figure 1: DTP start page

Figure 2: DTP project description

Deutsches Terminologie Portal - Microsoft Internet Explorer

Datei Bearbeiten Ansicht Favoriten Extras ?

Adresse http://www.iim.fh-koeln.de/dtp/dtp_frameset.html Wechseln zu

Deutsches Terminologie-Portal

Home • Kontakt • Feedback • Newsletter

- Texte zur Terminologie
- Literatur, Publikationen
- Terminologie-bestände
- Ausbildung an Hochschulen
- Veranstaltungs-kalender
- Werkzeuge
- Terminologie-normung
- Organisationen
- Projekte
- News
- Webterm
- Fakultät 03

Literatur und Veröffentlichungen

- Deutschsprachige Literatur (grundlegende Werke)
- Internationale Literatur (grundlegende Werke)
- Literaturdatenbank

Deutschsprachige Literatur (grundlegende Werke)

- **Arntz**, Reiner; **Picht**, Heribert; **Mayer**, Felix (2002): Einführung in die Terminologiearbeit. Hildesheim: Olms-Verlag [4. Auflage, frühere Auflagen nur Arntz/Picht]
- **Felber**, Helmut; **Budin**, Gerhard (1989): Terminologie in Theorie und Praxis. Tübingen: Narr Verlag
- **Hohnhold**, Ingo (1990): Übersetzungsorientierte Terminologiearbeit. Stuttgart: InTra [vergriffen und schwer zu beziehen]
- **Hoffmann**, Lothar; **Kalverkämper**, Hartwig; **Wiegand**, Herbert Ernst (Hrsg.) (1999): Fachsprachen: Ein internationales Handbuch zur Fachsprachenforschung

Start Kalen... HP Sc... 2 Ne... Micros... 2 Wi... Deut... 09:39

Figure 3: DTP terminological literature and bibliographical database

Deutsches Terminologie Portal - Microsoft Internet Explorer

Datei Bearbeiten Ansicht Favoriten Extras ?

Adresse http://www.iim.fh-koeln.de/dtp/dtp_frameset.html Wechseln zu

Deutsches Terminologie-Portal

Home • Kontakt • Feedback • Newsletter

- Texte zur Terminologie
- Literatur, Publikationen
- Terminologie-bestände
- Ausbildung an Hochschulen
- Veranstaltungs-kalender
- Werkzeuge
- Terminologie-normung
- Organisationen
- Projekte
- News
- Webterm
- Fakultät 03

Organisationen und Institutionen

- Organisationen im deutschsprachigen Raum
- Europäische Organisationen und Netzwerke
- Internationale Organisationen und Netzwerke

Organisationen im deutschsprachigen Raum

CiTaL
Centrum für Internationale
Terminologie und Angewandte Linguistik
Fachhochschule Konstanz
Reichenaustrasse 81c
D-78467 Konstanz
Tel.: +49 7531 9836 20
Fax: +49 7531 9836 18
http://cital.fh-konstanz.de
Email: CiTaL@fh-konstanz.de

DTT
Deutscher Terminologie-Tag
e.V.
c/o Jutta Zeumer, 1.
Vorsitzende
Hauptstr. 6
D-69117 Heidelberg
Tel.: +49 6221 27248
http://www.dttev.org
Email: JuttaZeumer@aol.de

Start Kalen... HP Sc... 2 Ne... Micros... 2 Wi... Deut... 09:42

Figure 4: DTP terminology organisations and institutions

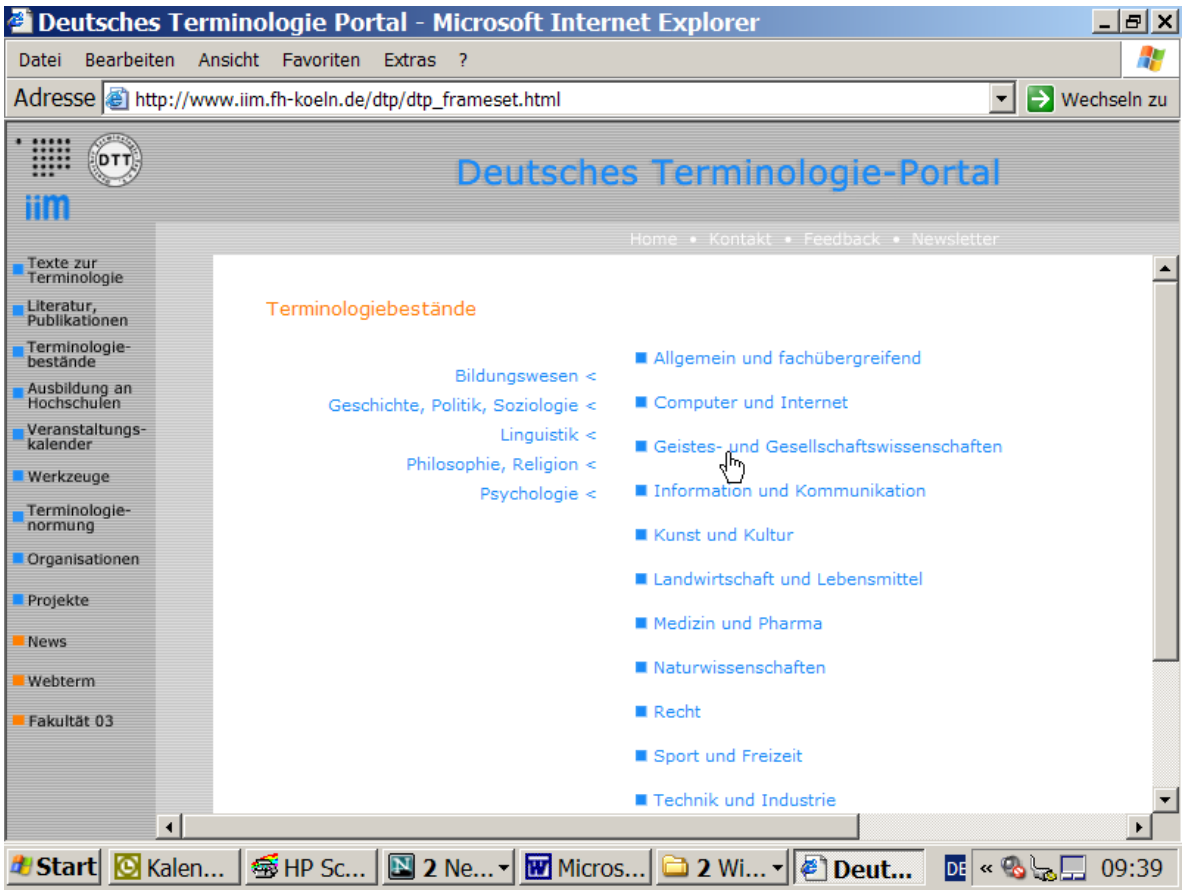

Figure 5: DTP terminology collections – subject field list

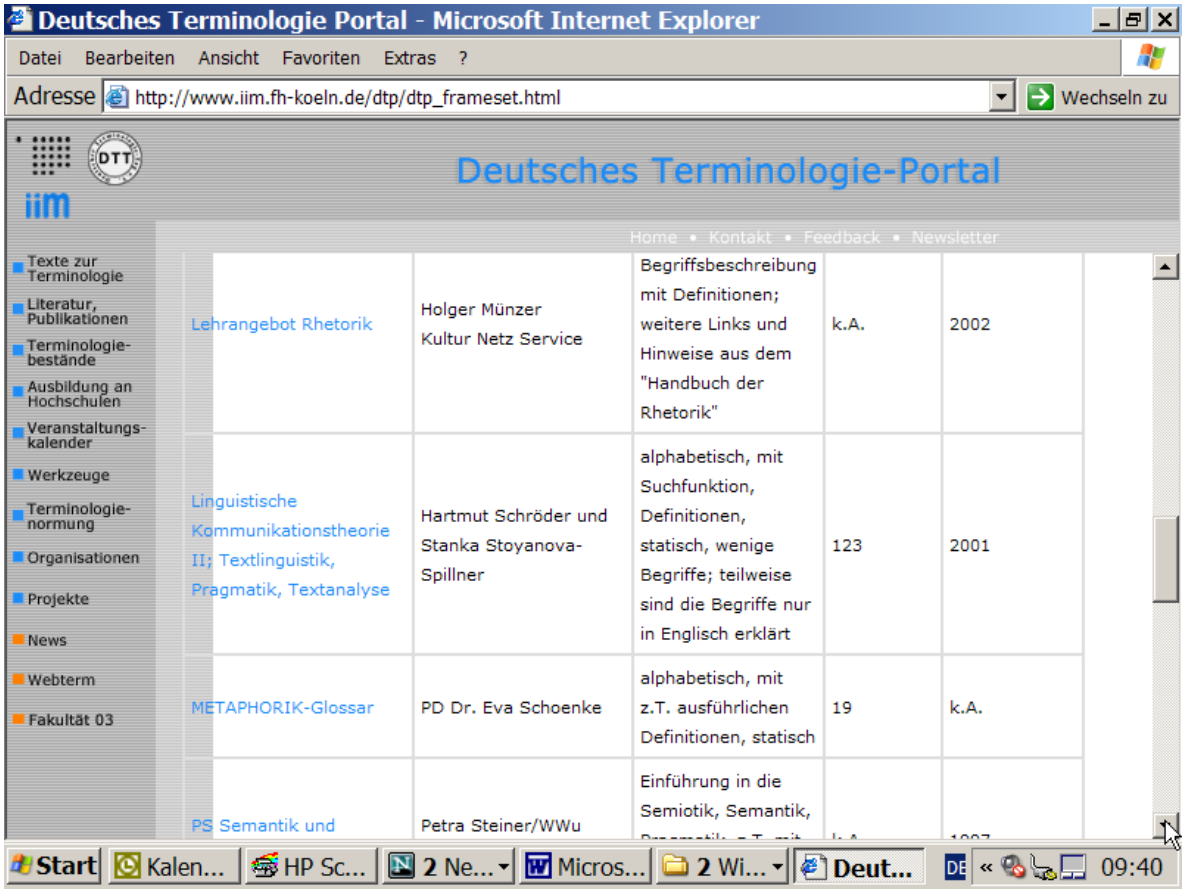

Figure 6: DTP terminology collections – descriptions and links

Deutsches Terminologie Portal - Microsoft Internet Explorer

Datei Bearbeiten Ansicht Favoriten Extras ?

Adresse http://www.iim.fh-koeln.de/dtp/dtp_frameset.html Wechseln zu

Deutsches Terminologie-Portal

Home • Kontakt • Feedback • Newsletter

- Texte zur Terminologie
- Literatur, Publikationen
- Terminologie-bestände
- Ausbildung an Hochschulen
- Veranstaltungs-kalender
- Werkzeuge
- Terminologie-normung
- Organisationen
- Projekte
- News
- Webterm
- Fakultät 03

AUSBILDUNG AN HS

Ausbildung in Deutschland >
Ausbildung in der Schweiz >
Ausbildung in Österreich >

Ausbildung in Deutschland (alphabetisch nach Ort)

- Hochschule **Anhalt** (FH), Studiengang Fachübersetzen
- Humboldt-Universität zu **Berlin**:
  - Anglistik und Amerikanistik
  - Romanistik
  - Institut für Slawistik/Russistik
- Universität **Erlangen-Nürnberg**, Institut für Fremdsprachen und Auslandskunde
- Fachhochschule **Flensburg**, Studiengang Techniküberset zen
- Institut für Übersetzen und Dolmetschen, **Heidelberg**
- Universität **Hildesheim**, Institut für Angewandte Sprachwissenschaft

Start Kalen... HP Sc... 2 Ne... Micros... 2 Wi... Deut... DE « 09:41

Figure 7: DTP terminology courses at universities – list of universities

Deutsches Terminologie Portal - Microsoft Internet Explorer

Datei Bearbeiten Ansicht Favoriten Extras ?

Adresse http://www.iim.fh-koeln.de/dtp/dtp_frameset.html Wechseln zu

Deutsches Terminologie-Portal

Home • Kontakt • Feedback • Newsletter

- Texte zur Terminologie
- Literatur, Publikationen
- Terminologie-bestände
- Ausbildung an Hochschulen
- Veranstaltungs-kalender
- Werkzeuge
- Terminologie-normung
- Organisationen
- Projekte
- News
- Webterm
- Fakultät 03

nach oben ^

INSTITUT	**Fachhochschule Köln** **Fakultät 03**
ANSCHRIFT	Mainzer Straße 5 D-50678 Köln
HOMEPAGE	http://www.spr.fh-koeln.de/
VERANTWORTLICHE PERSON(EN)	Name: Prof. Dr. Schmitz Vorname: Klaus-Dirk Email: klaus.schmitz@fh-koeln.de
SPRACHEN	Unterrichtssprache(n): Deutsch (Englisch, Französisch, Spanisch) Arbeitssprachen: Englisch, Französisch, Spanisch
ZWECK DER AUSBILDUNG	Studienmodul "Sprachdatenverarbeitung" (SDV) mit theoretischen und praktischen Veranstaltungen zur Terminologiearbeit (-methodik)
AUSRICHTUNG	Komponente des Diplom-Studiengangs
RAHMEN	8-sem. Studium zum Diplom-Übersetzer bzw. Diplom-

Start Kalen... HP Sc... 2 Ne... Micros... 2 Wi... Deut... DE « 09:41

Figure 8: DTP terminology courses at universities – course descriptions

Deutsches Terminologie Portal - Microsoft Internet Explorer

Datei Bearbeiten Ansicht Favoriten Extras ?

Adresse http://www.iim.fh-koeln.de/dtp/dtp_frameset.html Wechseln zu

Deutsches Terminologie-Portal

Home • Kontakt • Feedback • Newsletter

- Texte zur Terminologie
- Literatur, Publikationen
- Terminologie-bestände
- Ausbildung an Hochschulen
- Veranstaltungs-kalender
- Werkzeuge
- Terminologie-normung
- Organisationen
- Projekte
- News
- Webterm
- Fakultät 03

Werkzeuge

- Terminologieverwaltung
- Terminologieextraktion

Terminologieverwaltung

Werkzeug	Beschreibung	Kontakt
CATS - Computer-Aided Terminology System	sprachpaarbezogen, begriffsorientiert, vordefinierte Struktur, Verknüpfung der Einträge mit Bild- und Tondaten	CATS c/o Univ.-Prof. Dr. P. A. Schmitt Augustusplatz 10 / 11 D-04109 Leipzig http://cats-term.com
Lingo	Erstellung zweisprachiger Glossare, kompatibel mit allen Windows-Textverarbeitungssystemen, Toolbar-Oberfläche kann im Vordergrund des	Julia Emily Software Chambéry Frankreich E-Mail:

Start Kalen... HP Sc... 2 Ne... Micros... 2 Wi... Deut... DE « 09:42

Figure 9: DTP description of terminology management and extraction tools

Deutsches Terminologie Portal - Microsoft Internet Explorer

Datei Bearbeiten Ansicht Favoriten Extras ?

Adresse http://www.iim.fh-koeln.de/dtp/dtp_frameset.html Wechseln zu

Deutsches Terminologie-Portal

Home • Kontakt • Feedback • Newsletter

- Texte zur Terminologie
- Literatur, Publikationen
- Terminologie-bestände
- Ausbildung an Hochschulen
- Veranstaltungs-kalender
- Werkzeuge
- Terminologie-normung
- Organisationen
- Projekte
- News
- Webterm
- Fakultät 03

KONTAKT - IMPRESSUM - ANREISE

Institut für Informationsmanagement (R249)
Ubierring 40
D - 50678 Köln
Tel.: +49 221 8275-3941
Fax: +49 221 8275-3942
Email: iim@fh-koeln.de

Projektleitung
Prof. Dr. Klaus-Dirk Schmitz
Fakultät 03
Mainzer Str. 5
D - 50678 Köln
Tel.: +49 221 8275-3272
Fax: +49 221 8275-3991
Email: klaus.schmitz@fh-koeln.de
Homepage: http://www.sp.fh-koeln.de/Personen/Schmitz/schmitz_home.html

Start Kalen... HP Sc... 2 Ne... Micros... 2 Wi... Deut... DE « 09:44

Figure 10: DTP contact information

E-Learning in Music: Insights Gained from the Compilation of an Electronic Database on African Music Terms

Maria SMIT
Department of Music, University of Stellenbosch, SA

Abstract: This paper deals with the formative evaluation of an electronic multimedia database on Xhosa music terminology. A theoretical framework is presented within which the compilation of the database has been conceptualised and executed. The lexicographical insights draw mainly on the writings of Herbert Ernst Wiegand, and the publications of several scholars working on hypertext, multimedia, and computer-based instruction, are taken into account.

The database of Xhosa music terminology is discussed, and examples are shown. The types of information included are, apart from the lemma signs, the pronunciation, translation equivalents, etymological information, onomasiological and semasiological information, with special emphasis on describing the cultural context in which the terms are used to the potential users. Illustrations, photographs, video clips, sound recordings and animations are included where relevant, and links are made to related parts of the database.

Undergraduate students and high school pupils, who form part of the target users, evaluated the database. The purpose of the evaluation was to obtain information on the user-friendliness of the database, in order to improve it and to make certain general recommendations for similar terminological databases, intended for e-learning. Using Ripfel & Wiegand's (1988) suggestions, the evaluation by the subjects was done by means of questionnaires and protocols. The suggestions of other scholars were also taken into account, and an existing 'usability questionnaire' was adapted to serve the purposes. From the results obtained by means of the evaluation, certain insights are gained, and recommendations for the improvement of similar databases are made.

1. Introduction

One of the main problems in the study of music in South Africa, is the fact that information on and sound recordings of the traditional, indigenous musical cultures are available in university libraries, museums and other research institutions (in the form of dissertations, journal articles and field recordings on tapes), but often not accessible to a wider public. It is important that information on these traditional cultures, which have been marginalized during the apartheid era, should be readily available in order to promote renewed interest in them. One way of preserving and documenting indigenous cultural information might be by means of a multimedia database. Therefore, it was considered a worthwhile project to embark on the compilation of a database of African music, to be used by undergraduate students and high school pupils who study music.

E-learning has become fashionable,[1] and the use of technology, also in education, has mushroomed over the past ten years (Nielsen 1995: x). Many studies on multimedia and hypertext, also in lexicography, are now available (cf. e.g. Leech & Nesi 1999; Park & Etgen 2000; Storrer 2001), as well as studies on the planning and compilation of computer-based instruction (cf. e.g. Dunnagan & Christensen 2000), and the evaluation thereof (cf. e.g. Cradler 1994; Cradler & Bridgforth 1996; Hempel 1995; Kallinowski et al. *s.d.*; Sherry et al. 1997). Many scholars claim that, under certain conditions, computer-based instruction is an ideal medium (cf. e.g. Walters & Gardner 1990; Roschelle et al. 2000).[2] It is believed that computer-based instruction enhances 'idea organisation' (Nielsen 1995: 90), 'knowledge construction' (Jonassen et al. 2000: 108-113; Mayer 2001: 13),[3] and 'retention' (Mayer 2001: 72-75) and 'transfer' of knowledge (Mayer 2001: 75-78).

2. The multimedia project on African music terms

This paper focuses on the evaluation of an electronic database containing African musical terms. This terminological database has been developed in the Department of Music at Stellenbosch University, South Africa. The project investigated the possibilities of applying the German metalexicographer Herbert Ernst Wiegand's (e.g. 1998) theoretical assumptions, which have up to the present mainly been directed towards monolingual and translation dictionaries, to the terminology of African music. The evaluation of the database was done in order to (i) improve the user-friendliness of the database, and (ii) make more general suggestions for the types and presentation of data which could be included in electronic terminology databases to be used in e-learning.

The database in question is frame-based, drawing on suggestions in Konerding (1993) and Konerding & Wiegand (1994).[4] The first completed section of the database contains terminology of Xhosa music. There are 722 terms with definitions in the glossary. Types of information which are included are, amongst others: (i) the lemma sign, followed by the language of origin, and the pronunciation (in a sound clip); (ii) translation equivalents (if any) and names of other similar instruments; (iii) a short everyday description of the meaning of the term (included in the glossary); (iv) etymological information (if available); and (v) a fuller description of the object or event, with cultural information on its function and role within the culture. In addition,

---

[1] Cf. Cedefop (2002: 3) for a wide definition of e-learning as *'learning supported by information and communication technologies'*.

[2] According to Roschelle et al. (2000: 98-101), most major studies that investigated the question whether or not e-learning is beneficial, indicate that it is.

[3] This view of learning concurs with Wiegand's (1998: 170) 'actional-semantic' point of view on the use of dictionaries, namely, that a user *constructs* (or 'reconstructs') the meaning of lexical items by means of the data presented in a dictionary (or, as in this case, an electronic database).

[4] Cf., also, Woodhead's (1991: 37ff) discussion of frames and its use in hypermedia.

video clips, illustrations, photographs, animations and sound recordings are added where relevant, and links are supplied to the different parts of the database. (Some examples from the database will be demonstrated during the session.) The database is stored in ToolBook II on CD-ROM, but it is foreseen that once it is completed, it will also be made available online.

This Xhosa music section of the database has been evaluated for its user-friendliness by a number of high school pupils and undergraduate students, who form part of the target users. The investigation took into consideration the viewpoints of research on dictionary use by Ripfel & Wiegand (1988) and Wiegand (1998: 967-969, 996), using questionnaires and written protocols.[5] The evaluation was viewed as formative evaluation (Bresler 1994; Chou 1999) by means of which the database can be refined and improved, and an adapted 'usability questionnaire', based on one by Kalawsky (*s.d.*) of Loughborough University was used.

3. The usability questionnaire

The evaluation firstly contained questions about personal information of the subjects (for example, their prior experience with computers). Secondly, it requested the subjects to write a short essay and some shorter paragraphs on information contained in the database, in order to determine how retrievable the data are. These assignments differed in nature, in order to induce different user searches in various parts of the database, and with different data types. Thirdly, the questionnaire posed several questions on the functionality, the lay-out and the help functions of the database, its consistency in terms of user expectations and needs, and the overall impression that the database had on the subjects.

The results of this usability test are currently being analysed, and will be further discussed during the presentation. Aspects, which will in particular be looked at, are, for example, whether the types of information are sufficient to give users a clear idea of the meaning and the cultural context of the terms. Problems that were encountered will be mentioned, and suggestions that may have implications for electronic terminology databases for e-learning in general will be made.

References

Bloom, C., F. Linton and B. Bell. 1997. Using evaluation in the design of an intelligent tutoring system. *Journal of Interactive Learning Research* 8/2: 235-276.

Bresler, L. 1994. What Formative Research can do for Music Education: a Tool for Informed Change. *The Quarterly Journal of Music Teaching and Learning* 5/3: 11-24.

[5] Cf. also Bloom et al. (1997) for descriptions of evaluations of electronic instruction programmes.

Cedefop (European Centre for the Development of Vocational Training). 2002. *Users' views on e-learning*. (Cedefop Reference Studies 29). Luxembourg.

Chou, C. 1999. Developing CLUE: A Formative Evaluation System for Computer Network Learning Courseware. *Journal of Interactive Learning Research* 10/2: 179-193.

Cradler, J. 1994. *Summary of Current Research and Evaluation Findings on Technology in Education*. <http://majordomo.wested.org/techpolicy> (Last accessed on 11/02/2001)

Cradler, J. and E. Bridgforth. 1996. *Recent Research on the Effects of Technology on Teaching and Learning*. <http://majordomo.wested.org/techpolicy> (Last accessed on 02/03/2001)

Dunnagan, C.B. and D.L. Christensen. 2000. Static and Dynamic environments: the ambiguity of the problem. In J.M. Spector and T.M. Anderson (eds.): 61-78.

Hempel, C. 1995. Computergestütztes Musiklernen – Strategien, Möglichkeiten und Grenzen, dargestellt am Beispiel der Schulung des musikalischen Gehörs. In P. Becker et al. (eds.). *Zwischen Wissenschaft und Kunst*: 109-125. Mainz: Schott.

Jonassen, D.H., J. Hernandez-Serrano and I. Choi. 2000. Integrating constructivism and learning technologies. In J.M. Spector and T.M. Anderson (eds.): 103-128.

Kalawsky, R.S. (*s.d.*). *VR Usability Questionnaire: Issue 2, Loughborough University*. <http://www.avrrc.lboro.ac.uk/jtap305/reports> (Last accessed on 02/05/2002)

Kallinowski, F. et al. (*s.d.*) Implementation von CBT in den chirurgischen Unterricht. <http://link.springer-ny.com/link/service/books/3/540/148981/fpapers/mmlern_kallinowski.pdf> (Last accessed on 14/08/2002)

Konerding, K.-P. 1993. *Frames und lexikalisches Bedeutungswissen. Untersuchungen zur linguistischen Grundlegung einer Frametheorie und zu ihrer Anwendung in der Lexikographie*. (Reihe Germanistische Linguistik 142.) Tübingen: Max Niemeyer.

Konerding, K.-P. and H.E. Wiegand. 1994. Framebasierte Wörterbuchartikel: Zur Systematisierung der lexikographischen Präsentation des Bedeutungswissens zu Substantiven. *Lexicographica* 10: 11-170.

Leech, G. and H. Nesi. 1999. Moving towards perfection: the learners' (electronic) dictionary of the future. In T. Herbst and K. Popp (eds.). *The perfect learners' dictionary (?)*: 295-306. (Lexicographica Series Maior 95.) Tübingen: Max Niemeyer.

Mayer, R.E. 2001. *Multimedia learning*. Cambridge: CUP.

Nielsen, J. 1995. *Multimedia and Hypertext. The Internet and Beyond*. Cambridge: AP Professional.

Park, O.-C. and M.P. Etgen. 2000. Research-based principles for multimedia presentation. In J.M. Spector and T.M. Anderson (eds.): 197-212.

Ripfel, M. and H.E. Wiegand. 1988. Wörterbuchbenutzungsforschung. Ein kritischer Bericht. *Germanistische Linguistik: Studien zur neuhochdeutschen Lexikographie* 6: 490-520.

Roschelle, J.M., R.D. Pea, C.M. Hoadley, D.N. Gordin and B.M. Means. 2000. Changing how and what children learn in school with computer-based technologies. *The future of Children* 10/2: 76-101. <http://www.futureofchildren.org> (Last accessed on 14/08/2002)

Sherry, L., D. Lawyer-Brook and L. Black. 1997. Evaluation of the Boulder Valley Internet Project: A Theory-based Approach to Evaluation Design. *Journal of Interactive Learning Research* 8/2: 199-233.

Spector, J.M. and T.M. Anderson (eds.). 2000. *Integrated and holistic perspectives on learning, instruction and technology: understanding complexity*. Dordrecht: Kluwer.

Storrer, A. 2001. Digitale Wörterbücher als Hypertexte. In I. Lemberg, B. Schröder and A. Storrer (eds.). *Chancen und Perspektiven computergestützter Lexikographie*. *Hypertext, Internet und SGML / XML für die Produktion und Publikation digitaler Wörterbücher*: 53-69. Tübingen: Max Niemeyer.

Walters, J. and H. Gardner. 1990. *Domain projects as assessment vehicles in a computer-rich environment*. <http://www.edc.org/CCT/ccthome/reports/tr5.html> (Last accessed on 16/05/2001)

Wiegand, H.E. 1998. *Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie. 1. Teilband*. Berlin: De Gruyter.

Woodhead, N. 1991. *Hypertext and Hypermedia: Theory and Applications*. Wilmslow: Sigma Press.

Translation Technology and Workflow Procedures in Technical Documentation Management

Frieda STEURS
Department of Translators and Interpreters, Lessius Hogeschool, Antwerp, Belgium

1. Introduction

This paper will concentrate on the actual applications of translation technology and terminology management in industry, viz. in the management process of technical and commercial documentation. In Europe, companies are faced with a number of challenges concerning technical documentation and commercial leaflets. For documents accompanying products, the highest standards in quality management have to be applied. Due to the multilingual European society, the pressure on companies to produce the correct documentation in 15 to 20 languages, localised for the very diverse European market, is very demanding, time-consuming and expensive. Terminology management and translation tools may provide some solutions to these problems, provided that they are used in a proper way, and integrated with utmost care in the workflow process of creating and translating technical documentation.

We will first concentrate on the source text level, discussing some issues concerning source text quality, writing and authoring skills and means to control the source text creation. Then we will present a case study in which we will focus on workflow management in the creation of technical documentation and translation. Finally, we will try to come to some conclusions in relation to correct workflow management for technical writing and translation.

2. Source text control

In today's electronic age, almost everyone's life is touched by technology. But even though people use high-tech products on a daily basis, many do not know the first thing about how these complex appliances work. That is where technical communication comes in. Often referred to as technical writing, this profession is a vital link to the vast, ever-changing and frequently confusing world of technology. Without it, people would be inventing products that the mass public would not be able to use.

When the computer revolution started, many companies were relatively unprepared when it came to producing users' manuals for their products. More often than not, the item's creator would assume the technical writing duties, resulting in a guide that only people with advanced engineering degrees could understand. Luckily for everyone involved, things changed. Technical communication has become a legitimate career that requires specific skills to succeed, even if the general public typically has no idea it exists. The most basic definition of a technical writer is

someone who puts scientific and technical information into easily understood language. They prepare operating and maintenance manuals, catalogues, parts lists, assembly instructions, sales promotion materials and project proposals, among other things. They also plan and edit technical reports and oversee the preparation of illustrations, photographs, diagrams and charts. Communicating complex technical information to different audiences is a major challenge for the technical writer. This type of communication takes on a lot of different forms, including software development, online help and Web page development.

The job of the technical writer is thus a very complex one, but a crucial one to the company, although this is not always recognised. However, researchers such as Jo-Ann Hackos proved that a good and thorough user investigation is absolutely crucial for the production of a good manual. She pursues the idea of the 'minimal manual', which focuses on the writing of concise and consistent instructions and correct, clear-cut information. This can only be done when we know which user will read the manual, and how his/her expectancies or pre-knowledge of the tool or product are.

Two examples illustrate the importance of investigating the knowledge and behaviour of the user. The first case relates to the documentation for an airline-reservation system. User investigation led to a reduction of the original manual with 80%, leaving a small, elegant manual (that was thus only 20% of the original size) and resulting in 80% less phone calls to the helpdesk. The second case concerns the use of an operational manual in the field of parcel delivery where due to a confusing and not clearly written manual, an annual loss of 74 billion USD had to be accounted for.

In this domain, more research has to be done concerning the quality of the source text, and the skills and requirements for the technical writing. One interesting development here is the use of controlled language. A controlled language is a language in which terminology, syntax, and/or semantics are constrained. In some ways, it is analogous to a style guide used by editors and writers to achieve a clear and consistent style and terminology for a particular publication. Controlled languages have also been developed using languages other than English as the 'source' language, such as German, Swedish, French, Spanish, Greek, and Chinese. Controlled languages are used to enhance the clarity, usability, transferability, retrievability, extractability, and translatability of documents. This is achieved through increased terminological consistency and standardisation, generally simplified sentence structure, and standardised document format and layout. Controlled languages are particularly effective in commercial or industrial applications such as the authoring of user manuals or maintenance manuals, where large quantities of complex documents are generated and updated on a regular basis, and where terminology is domain specific. Controlled languages are also used in domains where documentation is traditionally highly complex, arcane, or poorly written, such as in government, finance, and law.

Looking at the complex problem of technical documentation, both from the point of view of creating the documents in the source language, and translating these documents into a multitude of languages, we would like to present a case study.

3. Case study: Integrated Language Services (ILS)

Integrated Language Services (ILS) is a linguistics department in a large printing company. Located near Antwerp in Belgium ILS was founded in 1999 to meet the growing demand of Blondé's customers for high quality content and increased consistency within multilingual documents. The staff at ILS are dealing with all language-related issues including copywriting, translation, Translation Memories and terminology management, remodelling and fine-tuning of language workflows in Pan-European communication projects. It may seem obvious that many printed publications need translation. However, language and text tend to be treated in a step-motherly way in a printing company, while the focus is normally placed on the quality of the materials used in the printing process, such as paper, ink, colour, etc. One has to safeguard the quality of the language used in the texts and to improve service to the customers.

The problems involved comprise different aspects:

- quality of the source text;
- updating of the content;
- quality of catalogues:
  - content validation and control;
  - localisation issues (different products and specifications for different markets);
  - quick updating of technical data;
  - time to market;
- quality of technical manuals;
- quality of owner manuals;
- quality of commercial publications.

In addition to these aspects that involve the source text, translation issues arise:

- quality of the target text;
- localisation for different markets;
- content validation.

In order to tackle these problems, a 'one source, many outputs' solution would be ideal. However, it has proven to be very difficult to reach this type of solution with customers one is not that familiar with, or where the mother company is very far away, as in this case for ILS, where most customers are Asian companies, who assign the

entire process of producing and printing manuals for the European market (sometimes up to 28 languages) to one company.

The main tasks of ILS can therefore be analysed as follows:

- coordination and streamlining of the translation process;
- creation, hosting and updating of the Translation Memories and terminology databases, online or offline, Web or desktop, to ensure consistent translations throughout various publications as well as the use of validated in-house customer-specific terminology;
- dealing with all editing, copywriting and localisation issues;
- content creation and delivery for multilingual websites, in cooperation with multimedia and knowledge departments;
- source text control and management.

3.1. Workflow procedures

The workflow procedures are designed in close cooperation with the customers, but the basic structure of the approach is as follows:

- Pre-study of existing documentation, including:
  - Style briefing to translators and reference material evaluated by the customer.
  - Terminology study of source texts and mapping of inconsistencies.
- Implementation of MAHT.
- Evaluation by the customer and internal evaluation.
- Continue full project execution.
- Follow up and periodic evaluation.

Many Asian Customers turn to this printing office to take care of all their publishing needs for the European market. A European market that is very confusing to them, due to the large amount of cultures and languages, and this involves quite some difficult localisation techniques.

The Asian source texts show a number of problems:

- Very often they are already translated in a “draft” version: Japenglese.
- They are very often available in unknown data formats only, incompatible to most Translation Memories and file formats for printing.
- The content of the text still needs adaptation to the European market.

In some cases, the whole source text needs revision in order to be suitable for translation. It is very hard to convince the company of this type of additional tasks, involving delays and more costs.

The text is not only often grammatically and semantically inconsistent or simply incorrect, but frequently the author is a person working in the technical staff, who may be sensitive about technical elements, but insensitive to market and customer orientation.

In addition, rewritten source texts need to be accompanied by comprehensive local market guidelines when sent out to translators. Even with technical information such as owner manuals, there is always the need for some level of localisation. For example, technical specifications may differ from one country to another, and even the market orientation of products may differ from one country to another.

Another complication is text expansion. There may not be enough space reserved in the Asian master document for expansion into European languages. This forces one to rebuild the master template and layout and to adapt it for the European market.

3.2. Knowledge management

Knowledge management is the key term covering this evolution. Knowledge exists on the levels of:

- textual info;
- graphical info;
- corporate identity info.

These types of information need to be reused in various output formats, including CD-ROMs, websites, media-neutral databases, printed matter, etc. It goes without saying that implementing knowledge management requires serious investment, training, and internal restructuring. However, this investment will be returned on a mid-term basis. This is done not only by a reduction in time to market, but also in a highly differentiated output: from printed documents over websites to CD-ROMS etc.

Additional advantages of knowledge management include:

- improved consistency;
- reduction of budget strains resulting from communication errors;
- process streamlining;
- customer binding;
- production cycle cutback;
- corporate identity enhancement;
- market positioning;
- media-neutral storage, allowing for future re-use, e.g. using Translation Memories.

While it may be hard to convince a customer at first of the benefits of knowledge management, it is even harder to convince customers of the necessity of terminology management and MAHT.

As an adequate budget is nearly always absent, and the benefits are not immediately visible, 'selling' Translation Memory and terminology database solutions is not always easy. Customers are primarily interested in cutting back translation costs, and are hesitant to invest in long-term solutions. In addition, not all documents are suitable for processing with translation software and care should be taken when offering this service to a customer.

4. Conclusion

In this paper we tried to highlight some of the procedures in technical document creation and translation that may improve the whole process:

- source text control;
- terminology management;
- translation management;
- content management;
- critical analysis of the needs of the user;
- workflow management.

A number of interesting industrial cases prove that this method of handling technical documentation may lead to a more successful process management and finally to a better output.

The Development of Terminologies in African Languages as a Key to Sustainable Human Development and Empowerment

Nonkosi TYOLWANA
Convenor: Terminology and Lexicography, IsiXhosa National Language Body

1. Introduction

In the process of colonialisation, speakers of African languages lost their self-confidence, sovereignty, self-respect and the power to participate actively and consciously in the development and promotion of their languages.

There is a great challenge for the coordination and management of terminology in African languages. It demands imagination, courage, perseverance and self-control. Language development is not about the delivery of language goods to passive citizenry, but rather about *involvement*, *exclusivity* and growing *empowerment* of the speakers. Sustained human development is development that regenerates environment rather than destroying it – it *empowers people rather than marginalizing them*. It gives priority to the poor, enlarging their choices and opportunities and providing for their participation in decisions affecting them.

The RDP seeks to promote development as a people-driven process. It recognises the aspirations and collective determination of the people of South Africa as the country's most important resource. The commitments are therefore the elimination of structures that perpetuate exploitation and impoverishment (Liebenberg & Stewart 1997: 123).

The aim of this paper is to encourage transparency in the interaction between the 'collectors' of terminologies (language workers) who visit local communities and the 'holders' or 'beneficiaries' of these terminologies.

2. Sustainable human development

Our notion of sustainable human development refers to a process that is human-centred, equitable, and socially and environmentally sustainable. Sustainable human development is understood as a means for the expansion of the choice of individuals in a society. There are five aspects that define sustainable human development. They are empowerment, cooperation, equity, scarcity and security (Isata 2000: 25).

2.1. Language as sustainable human development

Local communities are "knowledge rich, but economically poor". The search is therefore for a middle way in the development of the linkage between communities and the formal processes of terminology management and the formal processes of terminology management. This includes the development of clear strategies that are aimed at the promotion of terminologies for the African languages.

Closely connected with this is the increasing concern regarding the professional conduct of terminology researchers. In this regard, especially the tradition of always leaving local linguistic communities anonymous in publications of term lists or term banks – even if these communities have been a key in coining or divulging the key terms in specific languages – is worrisome.

Galinski & Wright (2001) provide a powerful proposal regarding the *protection of the intellectual property rights* when they say:

> Terminologists collect information which they reserve for their own use or share with others, either gratis or for a fee. Consequently most terminology collection can be viewed as compilations in the sense of a copyright law … The protection of authors' rights with respect to the transfer and reuse of terminological data or their use to prepare derivative works not only involves the application of the above-cited protection for intellectual property rights, but also for dreams.

The Portfolio Committee on Arts, Culture, Language, Science and Technology of the Parliament of South Africa is in the process of putting in place a legislation that is aimed at the protection of the intellectual property rights.

An important question is: "How do current linguists or terminologists unearth this knowledge so that it can shape the consciousness of language workers." Vilakazi (1998: 12) echoes the same sentiments when he says:

> European languages have had to develop new concepts, words and flexibility in order to be the means of communication for industrialization and scientific revolution. The recent development of history of Afrikaans is an example, which can now be used to teach physics, medicine, psychology and so on.
>
> The question now is who should be involved in the development of African languages. This task should not be left solely to university-certificated language experts. Every village can identify such individuals who are the custodians of excellence, beauty and pride of the African languages of this country. These women and men should be incorporated into the language committees of terminology experts, to work side by side with the university professors, researchers, terminologists, and language development in charting the formal development of African languages.

2.2. The role communities can play in language development

African languages remain a problem in the country's effort to revitalise local communities. It has for example been recognised that English is valuable for international communication, but it alienates locally. The key to promote and develop terminologies that influence language development lies in knowledge deeply rooted in local communities, the majority of whom are speakers in rural and farm areas.

Terminology knowledge is characterised by being embedded in the cultural web and the history of people, including their civilisation. Terminology development and innovation systems must therefore be sustained through active support of the linguistic communities who are the keepers of these terminologies, the custodians of their

languages, their ways of life, their social organisations and the environment in which they live. This means that the terminology of African languages resides within these communities.

Terminologists are expected to serve the public in areas of their profession. The issue at stake is what the best way is to honour our 'terminology holders' for their valuable contribution? Who should take the initiatives and how? Who should fund such facilitation? Who should look for funding of such initiatives?

3. Action programme

3.1. General recommendations

The biggest challenge is to bring terminologists / linguists close to the centre of this vital debate. Communication among stakeholders is in a number of cases constrained by educational restrictions in the transfer of knowledge. The lack of communication skills is hampering the participatory implementation of language development at local level.

The challenges and issues at stake include the following:

- There should be enhanced support for terminology activities at local level.
- An inventory should list all terminology stakeholders, including community-based projects.
- Government in cooperation with language training institutions including NGOs and CBOs should extend and improve their training facilities for human resource development.
- There should be recognition of women as repositories of terminologies pertaining to female-related terms and so on.
- Develop a strategy for educating local communities about the benefits of African languages.
- Develop sound and ethical protocols to record 'secret' and 'sacred' terminology, and guide the research in this regard (e.g. terms related to circumcision).

3.2. Specific recommendations

1. Establishment of Language Development Centres

- Envisaged to be the historically disadvantaged universities and their primary function should be to act as collective community administrative agencies.
- Be accountable to DACST, PanSALB, language training institutions, and language NGOs and CBOs.

Their main functions would be:

- To facilitate the process of collection, distribution and dissemination of terminology-relevant term resources.
- Create and promote public awareness campaigns for and mass participation in terminology collection.
- Develop a directory of terminology inventors.
- Audit and compile a record of different language communities found in various communities in South Africa.
- Assist the custodians of these terminologies with the intention of identifying those that can be developed into business enterprises, ultimately leading to job creation.
- Build out the capacity of language students and terminology research team members in interacting with the holders of terminology.
- Train language students in terminology research methods applicable to interacting with communities.
- Involve the NLBs and PLCs in the creation of a series of language community structures and committees, and provide these structures and committees with a fair amount of work.

The following structures are proposed:

2. The National Steering Committee

This will be a supreme decision-making structure, consisting of representatives from the entire range of stakeholders, and will be responsible for funding, policy and strategy.

3. The Working Committee

This will act as an executive structure of the Steering Committee and will consists of managers from DACST, PanSALB, language training institutions, and language NGOs and CBOs.

4. The Inter-Provincial Operating Structure

This will consist of representatives of universities and provincial languages departments for the coordination of operational matters in the provinces.

5. The Technical Committee

This will consist of experts in terminology with a mandate to debate issues revolving around terminology development, management and coordination, and to make recommendations to the Steering Committee. They will also facilitate partnerships with

beneficiaries of terminology, institutions such as HSRC, CSIR, Science, Technology, Economics, Medicine, Education, etc.

4. Conclusion

The greatest challenge is to convince linguists / language workers that the public needs to know or to be involved in what is happening behind their language development walls. Unfortunately, no research has been carried out in South Africa to determine the level of interaction that linguists need to adopt with the public in order to create public understanding and support for research in language development issues. In the United Kingdom, however, the *Science and Technology Select Committee* of the House of Lords engaged a group of consultants to investigate this issue extensively, and to provide recommendation (cf. http://w.w.w.publications.parliament.uk/pa/Id199900/ Idselect/Idsctech/38/3801.htm).

In principle, this debate shows that language, like any other player in the public arena, ignores public attitudes and values at its peril. I feel that, by declaring the values which underpin their work, and by negotiating the values and attitudes of the public, linguists or language workers are far more likely to command public and government support.

References

Galinski, C. and S.E. Wright. 2001. Intellectual Property Rights. Copyright and Terminology. In S.E. Wright and G. Budin (eds.). *Handbook of Terminology Management*. Amsterdam: John Benjamins.

Isata, S. 2000. *Good Governance: An imperative for Africa in the third millennium*. Hout Bay: Gariep Publishing Company.

Liebenberg, S. and P. Stewart. (eds.). 1997. *Participatory Development Management and the RDP*. Cape Town: Juta & Co. Ltd.

Report. 1998. *Public Hearings on Indigenous Knowledge Systems*. Portfolio Committee on Arts, Culture, Language, Science & Technology.

Vilakazi, H.W. 1998. In S. Seepe (ed.). *Black Perspectives on Tertiary Institutional Transformation*. Johannesburg: Vivlia Publishers.

Towards the Creation of a Dictionary Culture in South Africa

Michele VAN DER MERWE
Boland College Stellenbosch, SA

1. Introduction

When I was a little girl and asked my mother what the meaning or spelling of a word was, she would tell me to find the answer in the dictionary and would send me off to find one. Using a dictionary became a habit and second nature to me. My mother was introducing me to a 'dictionary culture'.

Hausmann (1989: 13) defines dictionary culture as the adaptation of society to lexicography, in contrast with the notion of user-friendliness where lexicography adapts to society. He points out that the terms 'dictionary culture' and 'user-friendliness' are in direct contrast with each other. Actually, these two terms are intertwined because of the interaction between society (or community) and lexicography. Different relationships exist between them – they are interdependent.

2. Evaluation of the current situation regarding a dictionary culture in South Africa

At present there is no such culture in South Africa. Culture is not a static concept, however. Like language, it is dynamic. For the past 50 years, as a result of technological progress, we have been able to find out what is happening in any part of the world and to witness other nations' behaviour in a way that our ancestors could not.

The fact that culture is not a static concept can be advantageous and very good news in this case, because it means that a dictionary culture can also be established in South Africa. It means that society's beliefs, values, perceptions and attitudes can change in such a way that the use of dictionaries becomes a second nature.

In order to cultivate and develop a dictionary culture in a country the existence of lexicography has to be recognised as a subject field. Lexicographical research has to lay the foundation for sound dictionary projects. Lexicographers have to be trained to compile user-friendly dictionaries that are theoretically sound. We need to increase the public's awareness of lexicographic issues.

Progress, however, has already been made on the lexicographical scene in South Africa. In 1993 M. Alberts conducted a viability study of establishing lexicographical units in South Africa. Since 1997 meetings have been held to plan the establishment of lexicography units for each of the official languages of South Africa. In 1999 units for all eleven languages were established and they have projects running at the moment. The *National Lexicography Units* (NLUs) fall under the jurisdiction of the *Pan South African Language Board* (PanSALB). The aim and purpose of these NLUs are to compile dictionaries. NLUs can work closely with language communities to ensure

that their lexicographical needs are met when dictionaries are compiled. Language communities can communicate their special needs for specific dictionaries to the NLUs.

The compilation of dictionaries, the development of terminology, and the development of literature play an important role in the revival of the indigenous languages. All these processes are taking off in South Africa and are in a position to grow and mature. According to D. van Schalkwyk it is the responsibility and a core brief of PanSALB to see to it that each language in South Africa has the opportunity to grow and develop into an instrument serving all communication needs on all levels.

One of the challenges facing the NLUs at the moment is a financial one. Finances are needed to train lexicographers of the units, especially the chief editors of the projects. NLUs need to have sound management and often lexicographers are not trained as managers. Sound management skills are crucial to the success of the lexicographical projects.

3. Different potential role players for the establishment of a dictionary culture

There are different role players who can contribute to establishing a dictionary culture in a country. Language communities have a definite responsibility in the creation of a dictionary culture, as they must give certain input on their needs. They can make their needs public, for example through the media. A few years ago somebody wrote a letter to the newspaper *Burger* asking for a good yet affordable Afrikaans pocket dictionary. The community had a need for a specific dictionary and formulated that need. On the other hand, if one were to talk to Afrikaans lexicographers, they would say that it was not economically viable to publish a pocket dictionary for Afrikaans, because of the size of the market. So the following questions need to be asked: What is the responsibility of dictionary publisher? Can it be expected of government to subsidise the publishing of dictionaries? And what is the responsibility of the language community involved? The Editor-in-Chief of the WAT, D. van Schalkwyk, said in a radio talk show in December 2002 that the French Academy receives 20 to 60 million French Francs annually for dictionary projects. Clearly, they have an established dictionary culture in France. Can the same happen in South Africa?

The dictionary culture of a community could also be influenced by the media, for example in the form of dictionary reviews. Dictionary critics should be well trained because they can fulfil a very important function in setting high standards for dictionaries and also educating the public about dictionaries. Research can be done in elaborating criteria for assessing dictionaries. Bergenholtz & Tarp (1995: 232) stress the importance of workshop reports, which contain criticism of the relevant dictionaries used. Such reviewing as part of the preliminary work on a new dictionary will be all the more valuable as it is prepared by users who, besides taking a particular interest in dictionaries, are also well acquainted with the language(s) and subject field(s) in question. The approach in existing LSP dictionary reviews closely

corresponds to their LGP counterparts.

Hausmann (1989: 14) stresses the importance of lexicographical research to ensure that dictionaries are made more accessible to the community and also user-friendlier. Sound metalexicographical research would encourage better practical lexicography. Ideally a balance between the creation of a dictionary culture and user-friendliness should be found. Certain lexicographical adaptations could be made to meet the community's lexicographical needs, but the community should also receive teaching in dictionary use in order to improve its reference skills. This can result in better dictionary use and dictionary compilation.

The *African Association for Lexicography* (AFRILEX) has been supplying lexicographers of the established NLUs with lexicographical training. According to P. Silva (PanSALB minutes, 1997) part of the mission statement of the *Dictionary Unit for South African English* (DSAE) is the preparation and provision of short courses in lexicographical theory and methodology. D. van Schalkwyk (PanSALB minutes, 1997) describes the secondary functions of the WAT as follows: *'to act as growth point and stimulus for lexicographic activities and meta-lexicographic reflection and to contribute to the establishment and development of South African lexicography in support of the aims of the Pansalb'*. It is clear that these dictionary units are committed to the establishment and nurturing of a dictionary culture.

It is very difficult to measure or determine a dictionary culture of a community or country. According to Hartmann (1986) the lexicography or language of a country depends not only on the external history of that country or language, but also on the internal cultural traditions that help shape successive generations of dictionary making. The fruits of labour of these role-players can shape a dictionary culture in the community, but the community's traditions of established dictionary use can also ensure a history of dictionary making, so making it an interdependent process. In Germany, for example, 80% of all households possess dictionaries. In contrast to that, and according to the WAT, in 1990 over 7.25 million learners did not have access to dictionaries in South Africa.

4. Dictionary culture and early education

If Hausmann's (1989: 13) definition of dictionary culture is considered, namely the adaptation of society to lexicography, it is clear that a culture of dictionary use can be taught in order to establish a pattern of frequent dictionary consultation. From a very early age children can be exposed to dictionary use and they can be educated to know that dictionaries exist and what the functions of dictionaries are. The establishment of a dictionary culture in a community includes the supplying of a specific lexicographic education.

In a brief survey it was found that for children under six years there are some excellent dictionaries available in South Africa. For the very young learners there are thesaurus-like dictionaries available with a few lemmas, all illustrated in colour. The

target market is the baby or toddler, starting to familiarise herself with her world by learning names of objects. For the growing toddler there are some more complex dictionaries available, some with a hundred new words to learn.

A fine example of a dictionary for children in the primary school is *My First Bilingual Dictionary* (Human & Rousseau). It contains about 1000 lemmas. All lemmas are supplied with translation equivalents. Both lemmas and translation equivalents are used in explanatory sentences in the source and target languages. Most of the lemmas are supplied with colour illustrations. These illustrations are practical and simple, and information is conveyed in a user-friendly way.

Children's dictionaries could play a vital role in the creation of a dictionary culture by forming positive perceptions and attitudes towards dictionaries in children at a very young age ensuring lifelong dictionary use and thus lifelong learning.

5. How to encourage a dictionary culture in a community

One of the functions of the NLUs is to have closer cooperation and better communication with the language community. In the case of Sesotho the dictionary unit had input from the community when it was decided what kind of dictionary should be compiled. Sesotho users opted for a 32-page dictionary with primary school mathematics terms. The English terms will be translated in Sesotho and also be explained. Each translation equivalent will be supplied with an explanation in Sesotho. The target group for this dictionary will be Sesotho learners from Grades 3 to 7.

The aim of this dictionary for Sesotho learners is to get them accustomed to dictionary consultation and for them to see the dictionary as an important aid in their studies. At first users can get used to a subject-related dictionary, then a bilingual dictionary and later a monolingual dictionary. The aim is to win over a generation of dictionary users. The dictionaries are literally growing with the learners. As the learners grow, the dictionaries become more specialised and comprehensive.

To ensure the growth of a dictionary culture there should be communication and consultation between language communities, provincial language boards and education departments. Teachers have an important role to play in establishing a dictionary culture in schools, because they have to educate learners to use dictionaries. They can encourage learners to make use of dictionaries by including the teaching of dictionary use in the school syllabus. Dictionaries and their use should be taught from primary school level to ensure that a dictionary culture is established during the forming years. An entire generation can be brought up with dictionaries.

6. A new culture of e-terminology and e-lexicography

In the information- and knowledge-based society of the 21st century a new interaction between community and electronic lexicography has been developed, thus giving a new dimension to dictionary culture. Electronic dictionaries are fast becoming part of the online culture of society.

The new dispensation of online dictionaries brought some changes to the world of lexicography and terminology as we got to know it in the 20th century. There is a new interaction between society and lexicography. People make use of online dictionaries that are just a mouse click away from them. Users can also play a part in compiling dictionaries by sending suggestions of words online. Dictionaries can be upgraded constantly in an electronic environment. Maybe the problem of ephemeras can be solved in electronic dictionaries. Explanatory dictionaries can consist of standard word lists that do not change and addenda with ephemeras that can change as words become part of the standard language or become obsolete in the language.

According to Alberts (2002: 92) the number of computing and communication systems is projected to grow continuously within the next few years. From a prospective point of view new applications of the National Term Bank will increase. The combination of knowledge transfer, technology and specific knowledge, with terminology, is expected to create a national information structure situated in the National Term Bank that will enable new applications and open a new set of specialised communications.

Electronic developments hold new challenges for lexicographers. Online dictionaries call for new data structures. For example, lexicographers have to rethink the access structures that are currently being used in desk dictionaries and make provision for new innovative structures to be used online. The influence of e-lexicography on society could be enormous and could play a major role in establishing a dictionary culture.

7. Conclusion

Establishing a dictionary culture in a country is about the interaction between lexicography and community. To achieve this the attitude of the language community as well as the commitment of lexicographers to publish user-friendly dictionaries is crucial. A balance is needed between the two. A dictionary culture is needed more than ever in South Africa, especially with illiteracy becoming an ever-bigger problem in our country. In a multilingual society like South Africa dictionary users could be empowered and empowerment will be beneficial for all-purpose communication situations.

References

Alberts, M. 2002. E-terminology. *Lexikos* 12: 90-104.

Bergenholtz, H. and S. Tarp. 1995. *Manual of Specialised Lexicography.* John Amsterdam: John Benjamins.

Hartmann, R.R.K. 1986. The Training and Professional Development of Lexicographers in the UK. In R. Ilson (ed.). *Lexicography. An emerging International Profession*: 82-92. Manchester: Manchester University Press.

Hausmann, F.J. 1989. Die gesellschaftlichen Aufgaben der Lexikografie in Geschichte und Gegenwart. In F.J. Hausmann et al. (eds.). *Wörterbücher / Dictionaries / Dictionnaires. An International Encyclopedia of Lexicography*: 1-17. Berlin: De Gruyter.

Pansalb. *Lexicographic Meetings of Lexicographic Units 1997-1999.*

A Spellchecker for Afrikaans, Based on Morphological Analysis

Gerhard B. VAN HUYSSTEEN° & Menno M. VAN ZAANEN‡
Potchefstroom University for Christian Higher Education, SA° & University of Amsterdam, The Netherlands‡ and Tilburg University, The Netherlands‡

1. Introduction

Existing commercially available spellcheckers for Afrikaans (i.e. *PUK / Microsoft Speltoetser*; *Pharos Speller* (and hyphenator); *Ispell vir Afrikaans*) still don't comply with the general desiderata for spellcheckers (i.e. user-friendly, technically elegant, and linguistically justified, that gives the user high recall, high precision, and adequate suggestions), especially with regard to their lack in efficiency to cope with the structure and nature of Afrikaans as a semi-agglutinative language that allows for productive compound formation. The project under discussion is aimed at the development of an enhanced spellchecker for Afrikaans, which will not only improve in functionality and performance on existing Afrikaans spellcheckers, but also reach the benchmarks set by other comparable projects (e.g. the SCARrie project;[1] see also Van Huyssteen 2002). The three main areas of improvement are coverage / recall (i.e. the recognition of valid words as valid, and of errors as errors), flagging / precision (i.e. the correct indication of (potential) errors), and suggestion adequacy (i.e. to make useful suggestions for correcting errors) (Paggio & Underwood 1995: 6-8). The project commenced in February 2002,[2] and a first version of the new spellchecker will be released in July 2003. However, the project will continue after that, until the set benchmark is arrived at.

We will commence by discussing the general architecture of the spellchecker under development. The main focus in this paper will be on the evolution of the architecture of the spellchecker and the problems we are currently experiencing. We will then briefly discuss one of the morphological analysis modules, indicating the implications of this module on the spellchecker's recall and precision. As there is

[1] The spellchecker developed in the SCARrie project is based on the design and architecture developed in the CORrie project, a project aimed at the development of a morphological-based spellchecker for Dutch (Vosse 1994). As the SCARrie project improved on the CORrie project, the results obtained in the SCARrie project are set as the benchmark for the current project.

[2] This project has its seat at the Potchefstroom University for CHE (South Africa). In South Africa, the fields of Natural Language Processing and Computational Linguistics are by and large unexplored territory. However, since 2001 the Potchefstroom University prioritised these fields as strategically important research terrains, and invested extensively in the enhancement of personnel and other resources to develop Human Language Technologies for some of the South African languages. As the IT department at the Potchefstroom University had already developed a spellchecker for Afrikaans (the *PUK / Microsoft Spellchecker*), the development of an enhanced spellchecker was considered an 'easy' project to get hands-on experience in the above-mentioned fields.

currently no final product or prototype available, results presented in this paper are preliminary and to a large extent inconclusive.

2. General architecture

The spellchecker under development consists mainly of two kinds of modules, viz look-up modules and rule-based morphological analysis modules. The reusability of and the possibility to upgrade / refine these modules are constantly kept in mind during the development and integration phases.

2.1. Look-up modules

In the architecture, two look-up modules are mainly employed, viz a simple lexicon look-up module, and an error-detection module.

1. Lexicon look-up module

Although the ideal was initially set to reduce the size of the lexicon dramatically, it was later decided to increase the lexicon for two reasons: (i) in order to gain a few percentage points on lexical recall, and (ii) to intercept problems that might arise during morphological analysis. An increased lexicon should, however, not be problematic, due to the size and speed of modern computers (the spellchecker is developed for Office XP, and it is considered a given that end-users who use Office XP will have rather powerful machines), as well as the efficiency of modern data structures. The lexicon of the spellchecker therefore currently consists of circa 250,000 items, including ordinary words, proper names, specialised vocabulary (e.g. chemistry and biology terms), frequent compounds, etc.

2. Error-detection module

Initially the architecture provided only for the lexicon look-up module, but, because of problems arising with the morphological analysis later in the spellchecking procedure, it was deemed necessary to add an error-detection module. This module consists of two parts: (i) a look-up section that contains a list of frequently misspelled words, and (ii) a section where errors are detected based on 4-gram analysis at grapheme level.

2.1. *Look-up section*: The lexicon of the look-up section is based on a relatively small corpus of errors, collected via a Web-based competition that was launched during the project (see http://www.puk.ac.za/lettere/spel2002), as well as from available electronic data (e.g. from e-mails, messages on the Potchefstroom University's bulletin board, etc.). When a string is found during the look-up procedure, the misspelled word with its correct form is sent directly to the suggestion module. This lexicon will be extended and updated as more data becomes available.

2.2. *4-gram analysis section*: The 4-gram analysis section is based on a list of valid 4-grams for Afrikaans. To identify valid 4-grams, a software application was developed in Java, which enabled us to extract n-grams (i.e. n-graphs) from a corpus. The software also has the function to differentiate between n-grams that occur at the beginning, the end, and in the middle of words. As hyphens are not treated as word boundary markers during tokenisation at the start of the spellchecking session, and as hyphens can be used quite frequently in Afrikaans compounding, we had to accommodate hyphens in our lists of 4-grams. This procedure resulted in a list of 129,374 valid 4-grams, which are stored as three different sections of the 4-gram list.

Although the effectiveness and/or efficiency of this error-detection module cannot be precisely measured at this stage of the project, we believe that early error-detection will save on processing time, and also result in higher suggestion adequacy. It will however be essential to thoroughly evaluate this module in the final stages of the project (i.e. when a prototype of the spellchecker is available). For example, we would certainly experiment with n-grams, deciding whether we should rather use 3-trigrams or 5-grams. It could also be interesting to compare the overall results of the spellchecker with, without, or even only with the error-detection module.

2.2. Rule-based morphological analysis modules

In order to handle morphologically complex words (like compounds, derivations, derived compounds, etc.) a few rule-based morphological analysis modules are built into the spellchecker. These modules include a stemmer, a word segmenter, and a module that combines word segments in different sequential strings.

After a string was not found in the lexicon look-up module, and was judged by the 4-gram analysis module to be a valid string, the string is sent to a stemmer. The stemming procedure will be discussed in more detail later (see § 3.). If the string is still not found, it is analysed by the rule-based word segmentation module, where a string is segmented according to the word segmentation principles of Afrikaans. After the string has been segmented, the lexicon look-up procedure is applied to all the resulting strings. If all forms are found, the spellchecking session ends; if not, and if a word consists of three or more segments, the segments are sent to the module that combines word segments in different sequential strings (i.e. the de-segmentation module). In this module the different segments are glued together step by step and in different combinations. Each of these combinations is looked up in the lexicon. The session ends if the string is found. If not, the string is sent to the suggestions module, where suggestions are made based on an 'edit distance' algorithm.

3. Module: stemmer

Stemming algorithms are usually employed in Information Retrieval (IR) environments (see Porter 1980), and the aim of such stemmers is therefore to lump together *'nonidentical words which refer to the same principal concept'*, irrespective of

whether the resulting stem is a *'linguistically correct lemma or root'* (Paice 1990). In our case (i.e. with regard to a spellchecker) exactly the opposite holds true, where a linguistically correct form is more important than the semantics of the resulting stem.

The stemming algorithm that we are developing is based by and large on the design of the *Porter Stemmer for Dutch* (PSD) (cf. Kraaij & Pohlmann 1994). For instance, like in PSD, our stemming procedure is mainly based on affix stripping, and also includes some special conditions to cover certain phenomena (like the DupV-procedure, whereby closed syllables are identified and the vowel is subsequently doubled – see Kraaij & Pohlmann 1994: 170-171). Similarly, we include a measure condition, as well as some clean-up rules to render valid stems.

However, considering the different aims of the two algorithms, our stemmer also differs in some ways from PSD. In PSD only *'derivational affixes which do not substantially affect the information conveyed by the term'* (Kraaij & Pohlmann 1994: 170) are removed, whereas in our case, we are not restricted in this way. For example, the prefix *on–* 'un–' should, for semantic reasons, not be removed in an IR environment, whereas it can be harmlessly stripped off in a spellchecker. This entails that we do not have to restrict our algorithm to only the most frequent affixes (like in PSD), but that we can also include less frequent affixes, resulting in a longer list of derivational affixes that are removed by our stemmer than that of PSD. However, in order to preserve the efficiency of the stemming algorithm, we constrain ourselves to only include less frequent affixes that are highly regular (i.e. that combine only with free stems, or with a limited number of bounded stems), and do not cause any over-stemming problems. With regard to derivational affixes, our stemmer covers all affixes that are traditionally considered to be derivational affixes (see Jenkinson 1993).

Another (slight) difference between the two stemmers has to do with the grouping and ordering of rules. In PSD, the stemming rules (including the clean-up rules) were clustered into six groups in order to accommodate the level at which the affixes occur in the word formation process (Kraaij & Pohlmann 1994: 170). In our stemmer, the rules are clustered in two main groups, viz. in an inflection stemmer, and a derivational stemmer. Within each of these two groups, rules are carefully grouped together and ordered according to their formal and functional behaviour (e.g. all the plural suffixes are grouped together, and the past participle prefixes / infixes are ordered according to the length of the string). This clustering allows us to use differentiated and apposite procedures for each of the categories: for example, the procedure that verifies and, if necessary, removes the "*d*" or "*t*" after the plural "*–e*" rule has fired (e.g. *gaste* 'guests' → *gast* → *gas* 'guest'), needs not to apply after the diminutive rules have fired.

Despite these ordering and clustering, we still run into a considerable amount of problems, especially with regard to over-stemming. For example, the algorithm states that, after removing the comparative "*–er*", also remove the potentially redundant "*t*" (e.g. *sagter* 'softer' → *sagt* → *sag* 'soft', where *sagt* is a bounded stem, originating

from Dutch). If this algorithm is applied consistently, a word like *briljanter* 'more brilliant' is wrongly reduced to **briljan*. To solve this problem it seems as if we have two options: either include bounded stems in a special lexicon (thus *sagt* will be found, and **briljan* not), or add a look-up procedure after each rule that fires (thus the "*t*" of *briljant* will not be removed, because the string will be validated after *–er* is removed).

Although the former seems at first like the better, more 'economic' option, it is also problematic. To compile a lexicon of bounded stems will be a very difficult, labour-intensive and time-consuming process. Moreover, even a carefully crafted lexicon of bounded stems can result in wrong judgements: if someone would type **sagties* (instead of *saggies* 'softly'), the resulting stem *sagt* will be judged a valid stem after the suffixes *–ie–s* are removed. Of course, the same problem will occur in the second option if one does not deal carefully with the linguistic reality. For instance, if one applies the "*d / t*" removal rule non-discretionary, **sagties* will also be judged a valid string. However, in Afrikaans the potentially redundant "*d / t*" does not occur with diminutive suffixes, and one could therefore specify that this rule should not apply after the diminutive rules have fired.[3] As it seems like the more linguistically justifiable option, we therefore include look-up procedures in the stemming algorithm, where necessary.

A major concern at this stage is what the influence of this stemming algorithm will be on the overall processing speed of the spellchecker. Given that one of our main aims is to improve lexical recall (and not to increase the speed of the spellchecker), we have decided to continue along these lines. If deemed necessary in the evaluation phase, adjustments will be made to this algorithm (e.g. by restricting the stemming to only certain affixes).

Currently, more than 200 rules are employed in the stemmer to handle a set of frequently (as well as less frequently) occurring inflectional and derivational affixes. Although our stemmer has not been tested and evaluated thoroughly and systematically (and neither has it been fine-tuned), it yielded 85% correct stems in a preliminary test conducted on a 1000-word sample. As an 85% success rate is probably not good enough for use in a spellchecker, further work will have to increase the performance and efficiency of the stemming algorithm.

4. Conclusion

Although we are at this stage of the project not able to reach final conclusions, it seems as if simple morphological decomposition could cause some unwanted analyses, which could lower the precision and the processing efficiency of the spellchecker. However, by introducing additional techniques and other measures, these problems could be minimised. For instance, instead of using word segmentation and documentation

[3] Of course, the *–ie* suffix in *saggies* is not a diminutive suffix, but it has the same form as the regular *–ie* diminutive suffix in words like *boekie* 'booklet' or *plekkie* 'small place'.

modules, one could rather introduce a 'longest string' match algorithm, or one could prevent over-stemming mistakes by limiting the stemming algorithm to only the most frequent affixes. We are also considering using a Part of Speech tagger to prevent mistypings like *dieman* 'the+man' to be analysed as a valid string during word segmentation. These and other techniques will be explored further in the current project.

Acknowledgements
We thank and acknowledge the following members of the research team for their inputs in this research: Roald Eiselen, Christo Els, Petri Jooste, Christo Muller, Sulene Pilon, Martin Puttkammer, Werner Ravyse, and Daan Wissing. We would also like to express our gratitude towards Attie de Lange, Ulrike Janke, Boeta Pretorius, and Elsa van Tonder for technical, administrative, and legal support. The Potchefstroom University for CHE also sponsors this project generously – our thanks to Frikkie van Niekerk for his support.

References

Jenkinson, A.G. 1993. Die probleem van fleksie en afleiding in Afrikaans [The problem of inflection and derivation in Afrikaans]. *South African Journal of Linguistics. Supplement* 18: 100-122.

Kraaij, W. and R. Pohlmann. 1994. Porter's stemming algorithm for Dutch. In L.G.M. Noordman and W.A.M. De Vroomen (eds.). *Informatiewetenschap 1994: Wetenschapelijke bijdragen aan de derde STINFON Conferentie*: 167-180.

Paggio, P. and N.L. Underwood. 1995. Validating the TEMAA LE evaluation methodology: a case study on Danish spelling checkers. *Natural Language Engineering* 1/1: 1-18.

Paice, C.D. 1990. Another stemmer. *ACM-SIGIR Forum* 24/3: 56-61.

Porter, M.F. 1980. An algorithm for suffix stripping. *Program* 14/3: 130-137.

Van Huyssteen, G.B. 2002. Desiderata of spellchecking / spell-checking / spell checking: towards an intelligent spellchecker for Afrikaans. Paper presented at the one-day symposium *Developing Spelling Checkers for South African Languages*, 14 March 2002, Potchefstroom University for CHE, Potchefstroom, South Africa.

Vosse, T.G. 1994. *The Word Connection: Grammar-Based Spelling Error Correction in Dutch*. Ph.D. thesis. Leiden: Rijksuniversiteit Leiden.

The Compilation of a Quadrilingual Explanatory Dictionary of Chemistry

W. VAN ZYL DE VILLIERS
The South African Academy of Science and Arts, SA

Abstract: This paper describes the background to a project for the compilation of a *Quadrilingual Explanatory Dictionary of Chemistry* (QEDC) and the structure that was developed for this dictionary. It then highlights a number of relevant developments in the area of science, technology, education and language in South Africa since initiation of the project. In conclusion, the progress to date and the present status of the programme are briefly described.

1. Background

A number of policy documents of the South African government during the political transformation in the period 1990-1995 emphasised the importance of human resource development, the growth of the economy, and the role that science and technology could be expected to play in these transformations. Especially relevant in this regard were the 'White Paper on Reconstruction and Development' in 1994 and South Africa's 'Green Paper on Science and Technology'. The latter was followed in 1996 by the 'White Paper on Science and Technology', issued by the *Department of Arts, Culture, Science and Technology* (DACST), which proposed a National System of Innovation as framework for the utilisation of science and technology (S&T) for sustainable economic growth, employment creation, equity through redress and social development. The policy initiatives foreseen in the 'White Paper on Science and Technology' included, amongst others, mechanisms for human resource development and capacity building, as well as the promotion of public awareness and understanding of S&T.

During the same period members of the Chemical Sciences Division of *Die Suid-Afrikaanse Akademie vir Wetenskap en Kuns* 'The South African Academy of Science and Arts' became aware of the problems experienced by students from disadvantaged educational backgrounds when studying the natural sciences at tertiary level. This was ascribed to the fact that, for the vast majority, the language of textbooks and instruction at secondary and tertiary level was their second or even third language. This led to the situation where, contrary to well-known best educational practices, conceptualisation of fundamental scientific principles had to take place through medium of the non-mother-tongue. Discussions with linguists and lexicographers at the departments of Afrikaans and African languages of the University of Pretoria led, in 1996, to the formulation of a project for the compilation of a multilingual explanatory chemistry dictionary. In view of the uniqueness of the envisaged product it was decided to limit the dictionary to the field of chemistry. The lessons learnt from the

exercise could then in future be utilised in the compilation of similar dictionaries in other subject fields.

2. Structure of the QEDC

It was decided to include, in addition to English and Afrikaans as the two main teaching languages at South African secondary schools and tertiary education institutions, Sepedi and isiZulu in the dictionary. According to the South African census of 1996 these two languages have the largest number of mother-tongue speakers of the Sotho and Nguni language groups, respectively. These languages also supply access to others in the two language families. Furthermore, extensive expertise in lexicography relating to Sepedi and isiZulu is available at the University of Pretoria. The target users of the envisaged QEDC are learners at senior secondary level (Grades 10-12) and students enrolled for a first degree or diploma.

Against this background A. Carstens, in two articles in 1997 and 1998, described a number of challenges in the compilation of a dictionary with a strong pedagogical approach, having components of both translatory and explanatory dictionaries, aimed at a heterogeneous user group regarding encyclopaedic and foreign language proficiency. This meant that certain concessions had to be made in the structure of the QEDC when compared to the two types of dictionaries. English will serve as the source language, but with all encyclopaedic and linguistic information supplied in all four languages. This will facilitate conceptualisation in the mother tongue while, at the same time, the user also gains knowledge on terminology in the language of instruction. Access to the other three languages will be supplied through reverse term lists placed after the primary list of lemmas.

Definitions are formulated in relatively uncomplicated form, complemented with examples of the concept being described. Extensive linguistic information is supplied (where applicable), e.g. parts of speech, labels (old term, trivial name, etc.), homonyms, synonyms, antonyms and derivatives. In addition to simple examples, chemical information such as symbols, formulae and related terms are also given. Maintaining a balance between (i) consistency and simplicity relating to the type of definition as well as the language and terms used, and (ii) correct scientific information transfer, posed interesting challenges.

3. Recent developments relevant to the aims of the QEDC

During 2000-2001 the National Advisory Council on Innovation and the National Science and Technology Forum performed an investigation into the interaction between economic growth, science, technology and human capital. In their report 'Science and Technology – Nourishing Growth and Development in South Africa' it is shown that, historically, technological innovation has been the primary drive behind the growth of the major world economies. Furthermore, there is a strong correlation between the availability of human capital, research and development (R&D), and

economic growth. In turn, the mathematical and natural sciences are central to the development of these human resources.

Recently the *Department of Arts, Culture, Science and Technology* released a national R&D strategy built on three pillars, namely innovation, human capital and transformation, and the creation of an effective government S&T system. The second of these ought to (i) increase the number of women and previously disadvantaged persons in the sciences, and (ii) maximise the pursuit of excellence in global terms. The strategy document stresses the fact that South Africa lags far behind the developed world as well as many developing countries as shown by a number of important indicators, for example:

- R&D expenditure as a percentage of gross domestic product (GDP): SA 0.69% vs. 1.49% for Australia, 2.15% for the OECD countries and 2.47% for South Korea.
- Researchers per 1000 of the workforce: SA 0.71 vs. Australia 4.84 and South Korea 2.77.

The R&D strategy sets a number of goals to be achieved by 2012, e.g. to increase the number of matriculants with university exemptions in mathematics and science from the present 3.4% to 7.5%, to increase the proportion of S&T tertiary students to 30% and the number of S&T practitioners to 1.1 per 1000 of the labour force, with a significantly higher proportion of women and blacks. Total spending on R&D must also be raised to at least 1% of GDP.

Another indication of South Africa's problems regarding skilled human resources is painfully demonstrated by our position on the World Competitiveness Scoreboard compiled annually by the Institute of Management Development in Switzerland. After having been number 43 out of 47 countries in 1999 and 2000, South Africa's position improved slightly to 42nd and 39th out of 49 countries in 2001 and 2002, respectively. The major contributing factor to this poor performance is the very low rating received for South Africa's people in terms of education and workforce skills composition.

On the positive side, the challenges associated with the South African science and technology skills base that are required to underpin the much-needed economic growth have been addressed through many government initiatives over the past number of years. The goals of the national R&D strategy have already been mentioned before. The National Plan for Higher Education, issued by the *Department of Education* in February 2001, sets out to achieve the following:

- increasing the participation rate in higher education from 15% to 20% of the 20-24 year age group over the next 10-15 years;
- shifting the balance in enrolments away from the humanities towards economic sciences and S&T to achieve the ratio 40:30:30 over the next 5-10 years; and

- improving representation of black and female students in the latter two areas and at post-graduate level.

In June 2001 a national strategy for mathematics, science and technology education was launched to raise participation and performance by historically disadvantaged learners in maths and physical science, and to enhance the teaching capacity and skills to deliver quality education in these areas.

As far as language and multilingualism in South Africa are concerned, quite a number of significant developments have taken place since the start of the QEDC project. In March 1996 DACST held a workshop on 'The Feasibility of Technical Language Development in the African Languages' in Pretoria. In his opening address Dr. B.S. Ngubane, Minister of Arts, Culture, Science and Technology, pointed out: *'Knowledge transfer can only take place if a person understands the concepts being conveyed to him'*. He stated that the immense need for training programmes implied the creation of terminology in the African languages. During the workshop the role played by language, the mother tongue and terminology in empowerment and development was stressed by many speakers.

The underlying principle in the *Department of Education*'s Language in Education Policy, issued in July 1997, is one of additive multilingualism through maintenance of the home language, complemented by the effective acquisition of additional languages. However, although one of the aims of the policy is *'to counter disadvantages resulting from different kinds of mismatches between home languages and languages of learning and teaching'*, the only statement in the policy on the latter issue is the following: *'The language(s) of learning and teaching in a public school must be (an) official language(s)'*!

In May 2000 a working group on values in education, appointed by the Minister of Education, recommended two values in the area of language, namely the importance of studying through the mother tongue and the fostering of multilingualism. In a speech at the launch of a multilingual, multimedia teaching project in Soweto on 5 September 2000 Prof. Asmal again emphasised the difficulties experienced by both teachers and learners when using non-familiar languages. This was also highlighted in a survey by MarkData, commissioned by the *Pan South African Language Board* (September 2000), where respondents expressed strong preference for learning through their mother tongue complemented with good teaching of or learning through another official language or English.

One of the greatest challenges relating to the future role of science and technology in South Africa's development is perhaps defined best in the following statement in the DACST document 'Technology and Knowledge – Synthesis Report of the National Research and Technology Audit' (December 1998): *'Given that a competent human resource base remains the primary necessary condition for an effective science and technology system, mathematics, technology, physical science*

and language abilities at secondary school level represent the Achilles' heel of the preparation of future generations of science, engineering and technology workers in general, and researchers in particular.'

Against this background the participants in the QEDC project are more convinced than ever that the dictionary has the potential to make a significant contribution to the development of South Africa's people and economy.

4. Progress to date and status of the project

In view of the educational aims and target users of the QEDC, the corpus to be included was defined as those terms and concepts to be encountered by a student up to the end of his/her first year of chemistry studies at a South African higher education institution. Chemistry departments were requested to submit lists in this regard, which were then collated and refined by a panel of chemists from *Die Suid-Afrikaanse Akademie vir Wetenskap en Kuns*, resulting in a list of approximately 500 lemmas.

Definitions obtained from a number of internationally published chemistry dictionaries were adapted to the defined level of linguistic and encyclopaedic complexity for the QEDC by the subject specialists, followed by lexicographical processing by A. Carstens of the Department of Afrikaans at the University of Pretoria (UP). The other fields in the database were jointly completed by lexicographers and chemists. This was followed by translation into Afrikaans, after which the focus shifted to the Sepedi and isiZulu parts of the dictionary, the responsibility of E. Taljard and R. Gauton, respectively, of the Department of African Languages at UP.

Preliminary work relating to the Sepedi and isiZulu components of the dictionary has already been published by Carstens (1998) and Taljard & Gauton (2001).

References

Carstens, A. 1997. Issues in the Planning of a Multilingual Explanatory Dictionary of Chemistry for South African Students. *Lexikos* 7: 1-24.

Carstens, A. 1998. On justifying the compilation of a multilingual explanatory dictionary of chemistry. *South African Journal of Linguistics* 16/1: 1-6.

Taljard, E. and R. Gauton. 2001. Supplying Syntactic Information in a Quadrilingual Explanatory Dictionary of Chemistry (English, Afrikaans, isiZulu, Sepedi): A Preliminary Investigation. *Lexikos* 11: 191-208.

A Terminology Development Initiative (Corpus Planning from Below) in a Dual-Medium Science Project of PRAESA

Zola WABABA, T. MBATHA & B. MAHLALELA
Project for the Study of Alternative Education in South Africa (PRAESA), University of Cape Town, SA

1. Introduction

The aim of our paper is to highlight the link between the acquisition of multilingual knowledge and technology transfer. The project that we will report on was based on premises contained in the South African Constitution of 1996 and the national language-in-education policy for schools (LiEP, 1997). Furthermore, our basis is the Revised National Curriculum for Natural Sciences and Technology (2002). The National Curriculum envisages a teaching and learning milieu which requires that people in South Africa operate with a variety of learning styles as well as with culturally influenced perspectives, which in our view include African languages. This corresponds with the theme of this conference. The project that was undertaken at PRAESA involved the teaching of content subjects, including science, maths, etc. The approach that was used was a dual-medium approach based on the fact that in the teaching and learning of these subjects, two languages were already being used.

In our approach we conducted both intervention and research work. Our findings were as follows:

- There was need for terminology development in isiXhosa in order to teach the content subjects effectively.
- There were greater challenges encountered when developing Xhosa terminology with regard to standardisation.
- There was also a lack of resources for teaching content subjects (e.g. bilingual posters, textbooks and science equipment).
- There were no adequately trained science and maths teachers who were also Xhosa speakers.

The dual-medium science project for the intermediate and senior phases (Grades 4-7) was always the core of the programme. The starting point for the dual-medium approach was the recognition that it provided a way out of the stigmatised practice of MTE and the calamity of English-medium teaching and learning in a context where English was like a foreign language to learners.

In the two programme schools, the dual-medium approach took into account the reality that in the teaching and learning of 'content subjects' such as science, two languages were already being used, albeit unofficially and unequally. English was the official language when we arrived. In content subjects, including science, English was the language of: (i) the textbook and worksheets; (ii) of all written work by teachers

and learners, whether on the board or in notebooks; (iii) of all written assignments, tests, and exams; (iv) of key terms introduced orally in lessons; (v) of certain classroom management functions to establish formal authority, such as for greetings (Good morning class! Good morning teacher!). IsiXhosa, on the other hand, was used orally in science lessons: (i) by the teacher to elaborate on and illustrate concepts, albeit imperfectly; (ii) by learners in group discussions; (iii) by learners in reporting back to the class; (iv) by learners when answering the teacher's questions; (v) for other classroom management purposes, such as permission to leave the room, prayers, etc. What we found, therefore, were two half-LoLTs: English was generated and received mainly in the written mode, but seldom orally; and isiXhosa was used only orally, never in writing.

2. Codeswitching as a mask

A corollary of the unequal role allocation of the two half-LoLTs was teachers' use of codeswitching and codemixing. The programme team found that the simultaneous oral use of isiXhosa and English in the science classroom was widespread, intuitive, unsystematic, at times resourceful and yet fraught with problems. Over time, the programme team realised that some (not all) teachers tended to mask their own lack of confidence of the subject, as use of the 'other tongue' would certainly discourage learners from asking questions. This became apparent when the programme team asked teachers to use isiXhosa more systematically for oral and written classroom work. It also became clear, although difficult to establish empirically, that some (not all) science teachers ostensibly teaching through the medium of English had a relatively low proficiency in English. This exacerbated the problem of (official) English-medium teaching, and contributed to the widespread use of codeswitching and codemixing by teachers. Systematising the use of codeswitching and extending the use of isiXhosa to the written domain thus presented complex challenges. Over time it became an approach that teachers accepted and adopted, with some successes, as they saw the benefits for teaching and learning.

3. Finding scientific terminology in isiXhosa

The success of the science terminology project depended on a multi-faceted approach that simultaneously addressed questions about the nature and bias of scientific knowledge, educators' conceptual knowledge base, bilingual teaching approaches, teachers' proficiency in English, integrated bilingual assessment and learning support materials, and the development of scientific terminology in isiXhosa. The aim of this paper is to show how scientific terminology development was addressed by both teachers and the DMS programme team.

Initially, teachers problematised this aspect, saying English had all the words and isiXhosa was not ready for science teaching. However, the programme team took as a point of departure the insight that language develops through use. A range of

strategies would and should be employed to bring isiXhosa to the position English currently enjoys. Lexical borrowing, new coinages, unearthing of old concepts – all these were employed to convince teachers of the viability of using isiXhosa to teach science.

The programme team forged links with the *National Terminology Services* of DACST in an effort to deepen and extend the team's insights into term elaboration, the compilation of lists of scientific terms, and the meeting point of language planning 'from below' with language planning 'from above' in the standardisation of terminology. Teachers, understandably, were frequently unsure about which terms to use in isiXhosa, ranging from everyday words to more specialist terms. Complex issues such as the degree of admissibility of English (e.g. should the correct translation for germs be the more traditional *iintsholongwane* or was the borrowed term *iigerms* acceptable?) and the accuracy of scientific register had to be addressed in numerous workshops and discussions. The result was the beginnings of an exciting yet painstaking corpus-planning initiative *in embryo*.

4. Terminology development

We developed a strategy which we perfected over time that seemed to work extremely well with teachers, namely that of:

- Planning our Terminology Development Workshop with teachers, with them doing demonstration lessons.
- From the demo lessons we excerpted difficult terminology and provided definitions from the source language, viz. English. For this we bought ourselves a range of good textbooks and encyclopaedias so that we could readily excerpt the required definitions.
- We discussed a list of possible equivalents if there were any in isiXhosa, or simply came up with new definitions / coinages. Decisions about borrowings / loan words; total embedding; transliteration or coining had to be made. These were usually accompanied with a lot of arguments / discussion.

In the end we had to agree on the most appropriate. It sometimes meant ending a workshop not agreeing / being positive about the suitability of a particular term; but the participants were allowed to widely consult with others. Be it science teachers from the higher grades, or science educators from other schools; or even language teachers who usually had a deeper understanding of the mechanics of the language. The value of cooperation with other stakeholders when it comes to such a valuable corpus development initiative cannot be overstated. The success of the science programme attracted other non-science teaching staff to our workshops who wanted to be part of the dynamic process of language planning from below. The timing of the workshops was such that it didn't interfere with their teaching time. They wanted to be part of the science team the following year.

We realised the need to produce comprehensible definitions other than simply providing translated concepts / terms. The value of providing a Xhosa definition as well as the context in which it is used cannot be overemphasised. This means that although there may be many varieties of terms in various dialects spread across some geographical regions, a fuller definition and a context are more appropriate because they provide an elaborate explanation and understanding of the concepts taught. Hence we recommend a dictionary with both definitions and relevant contexts in which a term or a concept is used. Some scientific words from a Grade 4 Natural Sciences textbook in English are treated in Table 1 as an example.

Table 1: Some scientific words from a Grade 4 Natural Sciences textbook

Chlorophyll	A chemical substance that gives plants their green colour and which helps plants absorb light energy from the sun
Intlaza, ikloroform	**Yi-khemikali enika izityalo umbala oluhlaza, ekwancedisa izityalo ukutsala amandla okukhanya kwelanga**
Cell	The smallest working unit of a living organism
Ukhozwana lobomi	**Into encinci ebangela ukusebenza kwizinto eziphilayo**

5. Creating bilingual LSMs

A related (and valid) reason teachers initially used to question the viability of the proposed dual-medium approach was the paucity of learning support materials (LSMs) in isiXhosa. But where are the books? Where are the worksheets? Accordingly, materials development became a related focus, particularly in the third year of the programme when teachers were fully 'on board'. In workshops the science teachers develop resource packs with science teachers, drawing on material from existing textbooks and other sources. To this end, teachers had to be taught the skill of selecting suitable textbooks.

This meant deconstruction of nation-building, idealism inherent in the 'rainbow nation' textbook that came out post-1994, and questioning such notions as 'universal science' contained in them. It also meant ditching those books that made no distinction between English as primary language and English as additional language (EAL). Unsurprisingly, the creation of LSMs for teaching science bilingually, along with the other aspects of the science project mentioned above, took their toll on the science trainers. The experience provided a salutary lesson that teachers not altogether confident in their subject could not be expected to generate LSMs in a language they had not been trained in. Care had to be taken not to entrench a spoon-feeding, dependency mindset. This proved to be extremely difficult because of the level of under-training and the inadequate grasp of subject content on the part of most of the teachers.

Table 2b shows a worksheet which was prepared for a dual-medium lesson adapted from an English-only textbook (Table 2a).

Table 2a: Original English-only double-page spread of a Grade 4 Science textbook (Pluddemann 2002)

Air 1. Moving air - wind Here is a picture of a windy day on the beach. Look carefully at the picture. How do you know that the wind is blowing? Why are people packing up to leave the beach? Why is the little one crying? Do you know that there is air all around us? We can't see it, smell it or touch it. But when the air moves we can feel it and what it does. Wind is moving air. Sometimes when the wind blows very gently we can't feel it or see that it is blowing. But when the wind blows strongly, we can see what it does. *Page 1*	*Is it windy today?* 1. Go outside and look around you. Is it windy today? Can you feel the wind on your face? 2. Write a list of all the things you saw that told you the wind is blowing. *A very strong wind* Sometimes we get very strong winds called tornadoes. The wind in a tornado blows round and round very fast. Here is a picture of Welkom after a tornado. The wind in the tornado was travelling at about 400km an hour and knocked down these houses. Can you believe that moving air can cause so much damage? *Page 2*

Table 2b: Adapted Xhosa / English bilingual worksheet (adaptation by Zanele Mbude, the science specialist on the team)

Air - umoya Umoya oshukumayo 1. Moving air - wind Lomfanekiso ngowemini enomoya elwandle. Qwalasela okwenzekisa: Wazi njani ukuba umoya (air) uyavuthuza? Kutheni abantu begogosha izinto zabo besimka nje? Lentombazana incinane ikhalela ntoni na? Uyazi ukuba umoya usoloko ukhona usijikilezile? Asikwazi ukuwubamba nokuwubona. Kodwa xa umoya ushukuma siyawuva sibone nezinto ozenzayo. Umoya oshukumayo kuthiwa esiNgesini yiwind, loonto ithetha umoya ovuthuzayo. *Page 1*	*Is it windy today?* 1. Go outside and look around you. Is it windy today? Can you feel the wind on your face? 2. Write a list of all the things you saw that told you the wind is blowing. Umsebenzi - classwork **Kawucinge iindlela umoya oluncedo ngazo- Air is useful** • Ebantwini - How air is useful to people. • Kwizilwanyana - How air is useful to animals. • Kwizityalo - How air is useful to plants. • Kwizithuthi: imoto, inqwelo moya - How air is useful to cars and aeroplanes. **Atshe aphela amtyotyombe eKraaifontein!!! - Air can be harmful** **Ncokolani ngendlela owenzeke ngayo lomlilo. Bhalani amanqaku abalulekileyo.** *Page 2*

Follow-up interviews conducted with 24 teachers in the two schools involved indicated that the use of isiXhosa science terminology was “a learning experience”. The teachers mentioned that although there were different terms suggested for the Xhosa / English equivalents, they eventually reached a consensus. The teachers expressed that it was possible to teach science better in isiXhosa because science is not only an English phenomenon. The teachers said that dual-medium instruction was better because:

- learners understand better if the teacher uses two languages (home and an additional language);
- learners learn to express themselves easily and understand the subject matter very well; that is, children gain more participation and participate freely in lessons;
- performances in class improved.

The dual-medium programme, with terminology development at its heart, should be extended to other schools and to higher levels of education. This recommendation should go hand in hand with the development of multilingual knowledge in order to enhance technology transfer.

The Development of Computer-aided Term Extraction Software

Xinli YU & Min SONG
Department of Terminology Standardisation, China National Institute of Standardisation (CNIS), Beijing, China

1. Introduction

With the rapid increase of technical terms, the Internet has become the largest, up-to-date and fastest developing information resource database. On the one hand, a lot of new terms are continuously emerging – how to extract them? On the other hand, many terms have been acquiring new meanings and concepts – how to find these new meanings and new concepts? In addition, how to use terms and links between terms to turn this disordered information on the Internet into ordered and valuable information? Such problems were placed on our work schedule.

Modern terminology work includes two parts: manual and computer-aided, and the latter's proportion is on the increase by the day. However, only semi-automatic candidate-term extraction has been accomplished, because the examination and normalisation of terms is carried out manually, and only scattered terms can be extracted, not the extraction of terminology. Therefore, the software to be developed is only playing a role in supplementing new terms based on existing concept systems and terminology of subject fields. As for concepts corresponding to terms (or meanings of terms), some leads can also be found from specialised corpora by such means as KWIC (keyword in context) studies. This will contribute to the description of new terms (composition of definitions).

Now the *China National Institute of Standardisation* is cooperating with the Peking University to develop computer-aided candidate-term extraction software, and decided to take information science and information technology as its pilot field for establishing the database. We will gradually expand to other fields after having obtained experience in completing this software.

2. Selection and delimitation of subject fields with their scope

We have first selected terms from the related fields of information science and information technology, as key objectives of study, to develop term extraction software and to establish a database. In these two fields academic advancement is active and new technologies with new products are appearing one after the other. New terms are emerging frequently and also the terminology standardisation should be conducted most urgently. Besides, many basic disciplines are crossed and mixed together in these two fields, and also their term types are diverse, so the software developed will possibly be of higher transplantation value.

Now the term 'IT', which is used in industry circles, includes the content of telecommunication and computer technology. Actually, most of its content acts merely as a tool of information technology. Consequently, information science is its main part, and information technology but a part. As to telecommunication and computer technology, they belong to specific special fields and are also two marginal subsystems.

In knowledge classification, interdisciplinary relations include that of basis and application, as well as methods and objects, etc. As for information science, it is the result of crossing and mixing together many disciplines, including mathematics, logic, linguistics, psychology, computer technology, operation research, graphics technology, communication, library science and management science, etc. as its mother disciplines, which provide the theories (concepts, models, systems, etc.) and methods (techniques, tools, etc) respectively. In addition, information technology has found its application in more and more fields, so we can find many sub-fields known as "application of information technology X in field X". No matter these differences, there is no need for distinctions too fine-grained.

3. Establishment of related concept systems and terminologies

Concept systems and terminologies are now established mainly manually. The so-called related concept systems and terminologies include a general framework of human knowledge and that of mother disciplines, but they belong to a marginal part. It should be indicated that various mother disciplines always provided sub-disciplines with terms, in particular, mother terms consisting of nuclei of compound terms, and there are different relations between new and old terms, for instance, some new terms were formed by combination, abridgement or compression of old terms, or they originated from borrowing or extension of old terms, created in reference to old terms.

As a disciplinary framework, the knowledge framework of information science should include its history of development, that is, including human data, such as names of persons, institutions, books and journals, companies, places, etc. They have certain relations with terms, e.g. persons' names are always entered as compound terms, such as "X's theorem", "X's formula", etc.

The concept systems and terminologies of information science and information technology themselves are nuclei of terminology databases which are very important in developing our software to create new terms. Joint efforts are needed by terminologists and subject field specialists; but standardised reference books, textbooks, subject field encyclopaedia and dictionaries completed by specialists from various related subject fields can also be used for reference in this part. Due to the fact that this huge new field has a lot of content from a lot of basic disciplines, to straighten out the main ideas and to organise them in accordance with the principles of organising information by means of subject indexing, multidimensional retrieval is needed.

Equal attention to Chinese and English should be paid in this work, for most Chinese technical terms have English origins and even the technical documentation in English was done earlier than that of Chinese.

4. Design of the software

The software is aimed at daily real-time extraction of candidate terms form non-processed special information resources on the Internet. In designing this software the specific characteristics of Chinese should be taken into account, for Chinese has for example less morphological changes than English, while there are problems of word segmentation, etc. Term extraction in Chinese is thus more complicated than in English. However, solving many problems, Chinese researchers have obtained a great number of results that can be used and applied in our work.

People can command only a limited number of words. To name a great number of things one has to rely on compounds composed of individual words. Most technical terms are made up of noun compounds. It is reported that in documents of higher information density, including academic papers or product specifications, over sixty per cent of the noun compounds are terms. Besides, terms in compounds are advantageous for classification. Because Chinese term compounds are basically of attributive structure the nucleus of a compound represents the superordinate concept of this category. When a premodifier is added, a subordinate concept is formed. Thus, a cluster of related terms around this nucleus is formed. Viewed from another angle, many terms are compressed definitions, such as the term 'tubercular meningitis' shows: this is an inflammation (pathologic property) which came from encephalitis (anatomical), and is caused by mycobacterium tuberculosis (cause of disease). In extracting such a compound, one not only extracts a compound term, but also a corresponding definition simultaneously.

The nucleus of this series of term compounds is the most important, we call it the 'mother term', which is always the core term of a discipline and our primary object of extraction. In our manually-made concept systems and terminology lists of a subject field, most mother terms of this field have been provided, only newly emerging mother terms should be extracted by means of the software. According to foreign experience, they can be obtained by means of a frequency-based method. For instance, when one compares a word's frequency in special documentation with that in ordinary documentation, a ratio value (called weirdness ratio) is obtained. If such a value is notably higher, this word can be singled out as a term.

On the other hand, a rule-based method is a very effective one, as for the work of real-time extraction of candidate terms from non-processed resources on the Internet a rule-based method is more appropriate than a statistical method. In fact, there exist a lot of obvious clues in Chinese technical documentation, such as noun compounds in headlines, keywords or compounds with English abbreviations, Greek characters, digital codes or special symbols, etc. Cues providing semantic information

are plentiful, for instance, synonyms, explanations or definitions can be found between brackets, dashes, or as notes. In addition, some commonly used sentence patterns, such as "X X + (be, i.e., as) + Y Y", were induced from Chinese technical documentation, and used as templates for seeking synonyms, explanations and definitions.

Another commonly used method is 'concordance analysis'. In fact, this is a kind of 'word for word indexing', i.e. each and every appearance of a word is indexed. Lots of meaningful information can be found in a truncated KWIC context, such as compounds composed of certain words, their explanations or definitions, as well as relative terms, etc.

In short, the study and research of Chinese technical documentation and our handmade concept systems and terminologies of subject fields will provide us with many applicable methods.

Let's now turn to the sensitivity and specificity of term extraction software. Sensitivity means that the software will easily extract terms, but possibly they are not terms (pseudo-positive); and specificity means the software will extract terms only, but possibly true terms will be left out (pseudo-negative). Generally speaking, when term extraction software has a high sensitivity, its specificity will be lower. It is difficult to satisfy both sides. As a saying from the field of information technology goes: "There is often an inverse proportion between complete rate and accuracy rate". With regard to our work, we select high sensitivity, emphasise the complete rate, and then sieve manually.

5. Establishment of a corpus

Recently, corpus-based terminology work has made great progress, and terms extracted from corpora have been used to populate terminological databases, to provide translated terms for subject fields and tools for information retrieval. In order to obtain experience, we decided to set up a professional corpus already in the current stage of developing the software.

A corpus is a structured set of texts, systematically organised in accordance with certain specific needs. Normally, it is required that a corpus is representative, large and balanced, but a more decisive factor is the pre-processing, so as to assign grammatical labels. Extraction of subgroups from ordinary corpora is also carried out abroad, but it requires a very large ordinary corpus fully covering the date in this aspect. However, according to foreign experience, a professional corpus could be quite small in comparison to an ordinary one.

Correspondence

Prof. Khurshid **Ahmad**
Head: Department of Computing
University of Surrey
Guildford
Surrey, GU 27 XH
United Kingdom

Tel.: +00 44 (0) 1483 689322
Fax: +00 44 (0) 1483 689385
E-mail: k.ahmad@surrey.ac.uk | k.ahmad@eim.surrey.ac.uk
WWW: http://www.computing.surrey.ac.uk

Prof. A. **Akinyemi**
Center for Educational Technology
Sultan Qaboos University
PO Box 39
Al Koudh, 123
Sultanate of Oman

Tel./Fax: 00968 535206
Cell: 00968 9091848
E-mail: akinyemi@squ.edu.om

Dr. Mariëtta **Alberts**
Focus Area Manager: Lexicography and Terminology Development
Pan South African Language Board (PanSALB)
Private Bag X08
Arcadia, 0007
South Africa

Tel.: +27 (0)12 341 9638
Cell: 083 306 9924
Fax: +27 (0)12 341 5938
E-mail: marietta@pansalb.org.za

Dr. Bassey E. **Antia**
Department of Languages & Linguistics
University of Maiduguri
NG Borno State
Nigeria

E-mail: bantia1@yahoo.co.uk

Dr. Anne-Marie **Beukes**
Head: Language Planning
National Language Service
Private Bag X195
Pretoria, 0001
South Africa

Tel.: +27 (0)12 337 8366
Fax: +27 (0)12 324 2119
E-mail: td16@dacst5.pwv.gov.za
WWW: http://www.dac.gov.za/

Ms. Claudia **Blaschke**
TRADOS GmbH
Christophstr. 7

D-70178 Stuttgart
Germany

Tel.: +49 711 16877 59
Fax: +49 711 16877 50
E-mail: claudia@trados.com |
claudia.blaschke@trados.com

Prof. Sonja E. **Bosch**
Department of African Languages
University of South Africa (UNISA)
PO Box 392
Pretoria, 0003
South Africa

Tel.: +27 12 429 8253
Fax: +27 12 429 3355
E-mail: boschse@unisa.ac.za

Mr. Mark D. **Childress**
Information Coordinator / Terminology
SAP AG
Neurottstrasse 16
D-69190 Walldorf
Germany

Tel.: +49 (6227) 741527
Fax: +49 (6227) 78 21190
E-mail: mark.childress@sap.com
WWW: http://www.sap.com

Prof. André **Clas**
Département de linguistique et de traduction
Université de Montréal
C.P. 6128, Succ. Centre-ville
Montréal (Québec), H3C 3J7
Canada

Tel.: (514) 343 6111 # 3998
Fax: (514) 343 2284
E-mail: andre.clas@umontreal.ca |
clasand@ere.umontreal.ca

Dr. Jennifer **DeCamp**
MITRE Corporation
MITRE MS W430
7600 Old Springhouse Road
McLean VA 22102
United States

Tel.: +1 703 883 6060
Fax: +1 703 883 6930
E-mail: jdecamp@mitre.org
WWW: http://flrc.mitre.org

Mr. Gilles-Maurice **de Schryver**
Residentie Wellington
F. Rooseveltlaan, 381
B-9000 Gent
Belgium

E-mail:
gillesmaurice.deschryver@rug.ac.be
WWW: http://www.up.ac.za/academic/libarts/afrilang/elcforall.htm
(Electronic Corpora for African-Language Linguistics)

Dr. Joe **Downing**
Department of Corporate Communications
Southern Methodist University
United States

E-mail: jdowning@mail.smu.edu

Dr. Ronald D. **Eckard**
Department of English
Western Kentucky University
United States

E-mail: ronald.eckard@wku.edu

Dr. Christian **Galinski**
TermNet, International Network for Terminology
Aichholzgasse 6/12
A-1120 Vienna
Austria

E-mail: christian.galinski@chello.at

Prof. Rachélle **Gauton**
Department of African Languages
University of Pretoria
Pretoria, 0002
South Africa

Tel.: +27 (0)12 420 3715 (W) | +27 (0)12 361 3355 (H)
Fax: +27 (0)12 420 3163
E-mail: rgauton@postino.up.ac.za

Mr. Ewald **Gehrmann**
COO
STAR Technology & Solutions
Schoenaicher Str. 19
D-71032 Boeblingen
Germany

Tel.: +49 7031 41092 10
Cell: +49 172 9005550
Fax: +49 7031 41092 70
E-mail: ewald.gehrmann@star-group.net
WWW: http://www.star-solutions.net

Mr. Johan **Geldenhuys**
Head of Documentation: Nedcor Limited
First Floor, Finance Place, Nedcor Sandton
135 Rivonia Road
Sandown, 2196
South Africa

PO Box 1144
Johannesburg, 2000
South Africa

Tel.: +27 (0)11 295 7163
Cell: 082 900 2027
Fax: +27 (0)11 294 7163
E-mail: johang@nedcor.com

Mr. David **Joffe**
DJ Software
PO Box 299
Wapadrand, 0050
South Africa

Cell: 082 922 9932
E-mail: david_joffe@absamail.co.za

Ms. Barbara I. **Karsch**
Terminologist / Internship Coordinator:
J.D. Edwards
One Technology Way
Denver, CO 80237
United States

Tel.: +1 303 334 1656
Fax: +1 303 334 1679
E-mail:
barbara_karsch@jdedwards.com

Ms. B. **Mahlalela**
Project for the Study of Alternative
Education in South Africa
(PRAESA)
Private Bag University of Cape Town
Rondebosch, Cape Town
South Africa

E-mail: bthusi@humanities.uct.ac.za

Ms. Nolwazi **Mbananga**
Medical Sociology
The Medical Research Council Pretoria
1 Soutpansberg Road
Pretoria
South Africa

Private Bag X385
Pretoria, 0001
South Africa

Tel.: +27 (0)12 339 8500
Fax: +27 (0)12 324 1695
E-mail: nolwazi.mbananga@mrc.ac.za

Ms. T. **Mbatha**
Project for the Study of Alternative
Education in South Africa
(PRAESA)
Private Bag University of Cape Town
Rondebosch, Cape Town
South Africa

E-mail: mbttha001@mail.uct.ac.za

Mr. T. Xolile **Mfaxa**
Terminology Coordination Section
National Language Service
Department of Arts & Culture
Private Bag X195
Pretoria, 0001
South Africa

Tel.: +27 (0)12 337 8345
Fax: +27 (0)12 324 2119
E-mail: vt03@dacst5.pwv.gov.za
WWW: http://www.dac.gov.za/

Prof. Sergey **Papaev**
The Russian Research Institute for Classification, Terminology and Information on Standardization and Quality (VNIIKI)
Granatnyi pereulok 4
103001 Moscow
Russia

E-mail: papaev@vniiki.ru

Dr. V.V. **Potapov**
Leningradskoye shosse, 112-1 corpus 3, 609
125445 Moscow
Russia

Tel.: 007 095 4596953
Fax: 007 095 2462807
E-mail: potapova@linguanet.ru

Prof. Rodmonga K. **Potapova**
Leningradskoye shosse, 112-1 corpus 3, 609
125445 Moscow
Russia

Tel.: 007 095 4596953
Fax: 007 095 2462807
E-mail: potapova@linguanet.ru

Prof. Laurette **Pretorius**
Department of Computer Science and Information Systems
University of South Africa (UNISA)
PO Box 392
Pretoria, 0003
South Africa

Tel.: +27 (0)12 429 6727
E-mail: pretol@unisa.ac.za

Prof. D.J. **Prinsloo**
Department of African Languages
University of Pretoria
Pretoria, 0002
South Africa

Tel.: +27 (0)12 420 2320
Fax: +27 (0)12 420 3163
E-mail: prinsloo@postino.up.ac.za
WWW: http://www.up.ac.za/academic/libarts/afrilang/elcforall.htm (Electronic Corpora for African-Language Linguistics)

Dr. Uwe **Quasthoff**
Computer Science Institute, NLP Department
Leipzig University
Augustusplatz 10/11
D-04109 Leipzig
Germany

E-mail: quasthoff@informatik.uni-leipzig.de

Prof. Justus C. **Roux**
Director: Research Unit for Experimental Phonology
University of Stellenbosch
Private Bag X1
Matieland, 7602
South Africa

Tel.: +27 (0)21 808 2017
Cell: 083 2888 602
Fax: +27 (0)21 808 3975
E-mail: jcr@sun.ac.za
WWW: http://www.ast.sun.ac.za (African Speech Technology) | http://www.sun.ac.za/nefus (Research Unit for Experimental Phonology)

Prof. Irina N. **Rozina**
Dneprovsky pereulok, dom 124, korpus 5, kv. 41
344065 Rostov-na-Donu
Russia

Tel.: 7 (8632) 582538
Cell: 7 (8632) 614793
E-mail: rozina@iubip.ru | rozin@orbita1.ru

Prof. Dr. Klaus-Dirk **Schmitz**
Fakultät 03
Fachhochschule Köln
Mainzer Straße 5
D-50678 Köln
Germany

Tel.: 0221/8275 3272 (W) | 06897/74373 (H)
Fax: 0221/8275 3991 (W) | 06897/74364 (H)
E-mail: klaus.schmitz@fh-koeln.de
WWW: http://www.spr.fh-koeln.de/Personen/Schmitz/schmitz_home.html

Dr. Maria **Smit**
Music Department
University of Stellenbosch
Private Bag X1
Matieland, 7602
South Africa

Tel.: +27 (0)21 8082364 (W) | +27 (0)887 0656 (H)
E-mail: msmit@sun.ac.za

Ms. Min **Song**
Department of Terminology Standardisation
China National Institute of Standardisation (CNIS)
No. 3 Yuhui South Road
Chaoyang District
Beijing, 100029
China

Fax: +86 10 6492 1032 | +86 10 64921 8969
E-mail: songm@cnis.gov.cn

Prof. Dr. Frieda **Steurs**
Head: Research Centre Language and Computing
Department of Translators and Interpreters
Lessius Hogeschool
Sint-Andriesstraat 2
B-2000 Antwerp
Belgium

Tel.: +32 (0)3 2060 491
E-mail: frieda.steurs@lessius-ho.be

Dr. Elsabé **Taljard**
Department of African Languages
University of Pretoria
Pretoria, 0002
South Africa

Tel.: +27 (0)12 420 2494 (W) | +27 (0)12 332 1357 (H)
Cell: 082 353 6906
Fax: +27 (0)12 420 3163
E-mail: etaljard@postino.up.ac.za

Ms. Nonkosi **Tyolwana**
Language Services
RSA Parliament
PO Box 15
Cape Town, 8000
South Africa

Tel.: +27 (0)21 403 2777/8
Cell: 082 545 1051
Fax: +27 (0)21 462 1749
E-mail: ntyolwana@parliament.gov.za (W) | nonkosi@netactive.co.za (H)

Dr. Michele **van der Merwe**
Treurnichstraat 3
Paarl, 7646
South Africa

Tel.: +27 (0)21 8873027 (W) | +27 (0)21 863 1336 (H)
Cell: 083 2313613
Fax: +27 (0)21 863 1336
E-mail: michele@stelkol.sun.ac.za

Dr. Gerhard B. **van Huyssteen**
School for Languages
Potchefstroom University for CHE
Potchefstroom, 2531
South Africa

Tel.: +27 (0)18 299 1488
Fax: +27 (0)18 299 1562
E-mail: afngbvh@puknet.puk.ac.za

Ms. Linda **van Huyssteen**
Department of African Languages
University of South Africa (UNISA)
PO Box 392
Pretoria, 0003
South Africa

Tel.: +27 (0)12 429 8258 (W) | +27 (0)12 662 0145 (H)
Cell: 07 222 97 303
Fax: +27 (0)12 429 3221

E-mail: vhuysl@unisa.ac.za

Dr. Menno M. **van Zaanen**
ILK / Computational Linguistics, room R136
Tilburg University
PO Box 90153
5000 LE Tilburg
The Netherlands

E-mail: mvzaanen@uvt.nl

Dr. W. **van Zyl de Villiers**
Die Suid-Afrikaanse Akademie vir Wetenskap en Kuns
Private Bag X11
Arcadia, 0007
South Africa

Tel.: +27 (0)12 305 5630
Cell: 82 907 1123
E-mail: vzdevill@aec.co.za | vzdev@mweb.co.za

Mr. Zola **Wababa**
Project for the Study of Alternative Education in South Africa (PRAESA)
Private Bag University of Cape Town
Rondebosch, Cape Town
South Africa

Tel.: (012) 6504013
E-mail: zwababa@beattie.uct.ac.za | zwababa@humanities.uct.ac.za

Dr. Christian **Wolff**
Computer Science Institute, NLP Department
Leipzig University
Augustusplatz 10/11
04109 Leipzig
Germany

E-mail: wolff@informatik.uni-leipzig.de

Ms. Xinli **Yu**
Director: Department of Terminology Standardisation
China National Institute of Standardisation (CNIS)
No. 3 Yuhui South Road
Chaoyang District
Beijing, 100029
China

Fax: +86 10 6492 1032 | +86 10 64921 8969
E-mail: yux@cnis.gov.cn

SPONSORS

Department of Arts and Culture

International Network for Terminology

Transnet Limited

STAR Technology & Solutions

CONTRIBUTORS

Pan South African Language Board

John Benjamins Publishing

African Association for Lexicography