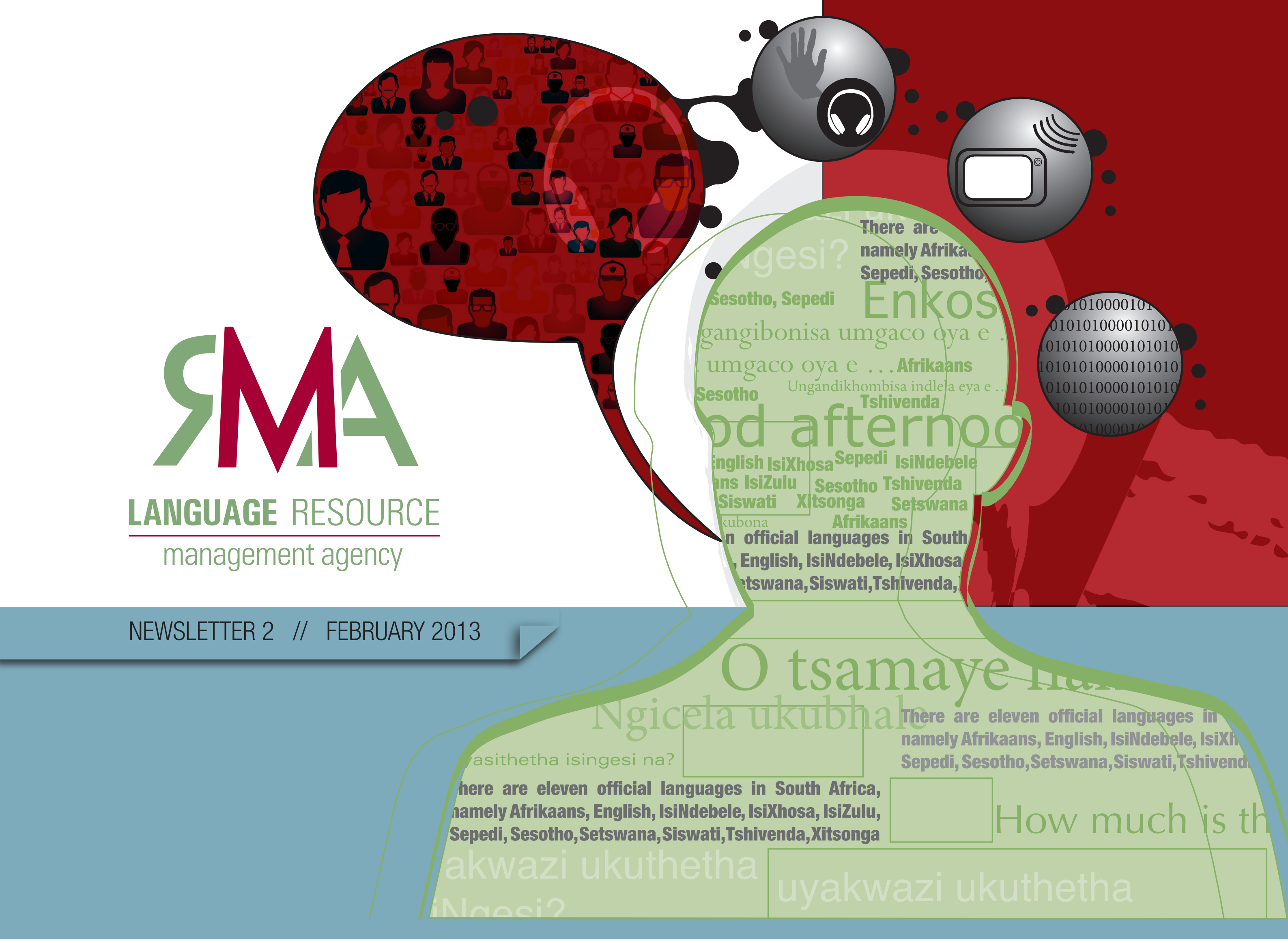

There are eleven official languages in South Africa, namely Afrikaans, English, IsiNdebele, IsiXhosa, IsiZulu, Sepedi, Sesotho, Setswana, Siswati, Tshivenda, Xitsonga
Enkos
Sesotho, Sepedi
gangibonisa umgaco oya e ..
umgaco oya e ... Afrikaans
Ungandikhombisa indlela eya e ..
Tshivenda
od afternoo
English IsiXhosa Sepedi IsiNdebele
IsiZulu Sesotho Tshivenda
Siswati Xitsonga Setswana
Afrikaans
O tsamaye
Ngicela ukubhale
asithetha isingesi na?
How much is th
akwazi ukuthetha
uyakwazi ukuthetha
iNgesi?
1010000101010

RMA

LANGUAGE RESOURCE
management agency

NEWSLETTER 2 // FEBRUARY 2013

Read about:

Finally a reality!

On 28 November 2012, CTexT® hosted the official launch of the long awaited South African Language Resource Management Agency (RMA) in Pretoria.

Prof Justus Roux of the NWU orientated the audience by providing an overview of language technology and what an RMA could mean for the South African context. International guest speaker, Dr Khalid Choukri, shed further light on the idea of an agency for the management and distribution of reusable digital text and speech resources speaking from a European point of view. As the general secretary of ELRA (European Language Resources Association) and CEO of its 'practical arm' ELDA (Evaluations and Language resources Distribution Agency), he provided the audience with an informative overview and statistics.

Dr Mbulelo Jokweni of the South African Department of Arts and Culture (DAC) reconfirmed the department's support of the RMA in promoting functional multilingualism in SA. Prof Etienne Barnard, also from the NWU, acted as Master of Ceremonies at the occasion.

The launch was well attended and comprised a diverse audience, including guests from the South African Translators Institute (SATI), Microsoft, the CSIR and Google.

Aspects of a legal framework
for language resource management

As can be expected, a data hub such as the RMA, where data is contributed and shared, does not operate without its own set of complex legalities.

The following aspects of a legal framework (based on Grover, Nieman, Roux & Van Huyssteen, 2012[i]) should provide some insight.

Stakeholders

The legal framework should be clearly defined by identifying the priority relationships of an RMA and by formalising these relationships through stakeholder analysis. Stakeholders may include primary and secondary content providers, service providers, end-users and networks. The important thing here is getting a full overview of existing relationships.

Language resources

It is important to identify priority language resources (LRs) that are (or should be) protected by legal means. Definitions for human language technology objects, such as 'corpora', 'lexica' and 'databases' should be carefully considered within a legal context, and an updated and comprehensive intellectual property (IP) register is vital.

Legal framework

The most important legal rights that come into play with respect to the provision of content to the RMA include privacy rights and IP rights.

The RMA should be aware of a general right to privacy associated with the content they distribute, to avoid infringement liability.

Furthermore, although language *per se* is not subject to IP protection, most of the LRs and associated technologies are governed by various restrictions. "A substantial amount of content could be exploited by an RMA because it belongs to the public domain. In Africa, traditional and indigenous knowledge and traditional cultural expressions or folklore do not fit easily into existing IP systems. In 2004, South Africa adopted an indigenous knowledge systems ("IKS") policy, which attempts to find a balance between respecting and protecting tradition on the one hand and enabling community economic development through commercial use on the other," says Grover *et al*., 2012.

There is also the open source IP domain to consider. Several external legal instruments (such as laws, treaties, conventions, etc.) exist that affect the open source IP domain. An RMA will have to decide which instruments are most important for its legal framework. A legal framework of an RMA also implies various internal legal instruments, such as end-user license agreements (EULAs), terms of references (TORs), service-level agreements (SLAs), etc. The South African (Language) Resource Management Agency will take care to ensure that data, both incoming and outgoing, adhere to the necessary standards and that the correct authorisations and license agreements are in place. Data will be handled with the utmost confidentiality (and be subject to anonymisation) so as to protect privacy at all times.

[i] Aditi Sharma Grover (1), Annamart Nieman (2), Justus C Roux (3), Gerhard B van Huyssteen (3)
(1) Human Language Technology Research Group, CSIR-Meraka Institute, Pretoria, South Africa
(2) Advocate, Member of the Johannesburg Bar, Sandton, South Africa
(3) Centre for Text Technology (CTexT®), North-West University, Potchefstroom, South Africa
[*In*: proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), Istanbul]

Prof Justus Roux (NWU)

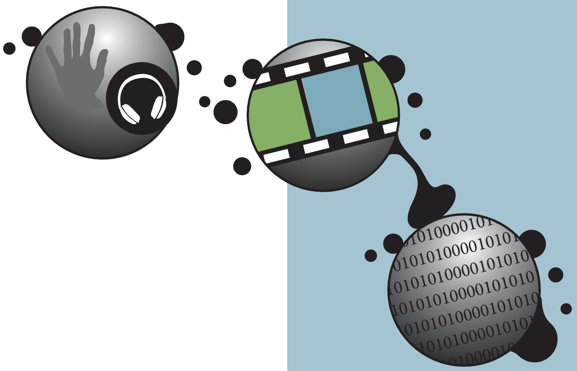
Dr Mbulelo Jokweni (DAC)

Funded by the Department of Arts and Culture

arts and culture
Department:
Arts and Culture
REPUBLIC OF SOUTH AFRICA

Managed by the Centre for Text Technology (CTexT®) of the North-West University (Potchefstroom Campus)

NORTH-WEST UNIVERSITY®
YUNIBESITI YA BOKONE-BOPHIRIMA
NOORDWES-UNIVERSITEIT
POTCHEFSTROOM CAMPUS



TST-Centrale: Trusted friend and ally

The Centre for Text Technology (CTexT®) at the North-West University has been appointed to set up the RMA in close cooperation with the Dutch HLT Agency (or TST-Centrale) over the next three years.

Remco van Veenendaal from the TST-Centrale visited CTexT in 2012 and was able to provide the RMA project team with valuable recommendations and suggestions. Therefore, as a valued collaborator and advisor to the setting up of the RMA, the TST-Centrale deserves a closer look.

According to Van Veenendaal, the TST-Centrale is the Dutch-Flemish Human Language Technology Agency for the management, maintenance, distribution and support of digital Dutch language resources. Most resources are the result of government-funded research programmes or projects. The resources are made available for use in research, education and (commercial) applications. The TST-Centrale is an initiative of the Dutch Language Union. "We collect, manage, maintain, distribute and support digital Dutch language resources. We are also the Dutch Language Union's CLARIN[i] Centre in the European CLARIN infrastructure," says Van Veenendaal.

Van Veenendaal suggests that "applications of language and speech technology play an increasingly important role in everyday life, and should be available in Dutch. For the research, education and development of applications, building blocks are required – language resources". According to him, the Dutch and Flemish governments have invested in the creation of language resources, resulting in text and speech corpora, monolingual and bilingual lexica and tools. The HLT Agency manages, maintains, distributes and supports these language resources.

"We are very happy to work with the RMA for a number of reasons. The cooperation strengthens the existing HLT links between the Netherlands, Flanders and South Africa. The RMA is also another example of having a central (government-funded) agency where a country's language resources are made available and kept available. Linking these initiatives and learning from each other's experiences are mutually beneficial. And last but not least: from our initial talks at the conference LREC 2010 onward, the contacts between the people involved were positive and, to paraphrase a saying I remember from an event about data centres, open (here: informal) when possible, closed (here: formal) when needed," said Van Veenendaal.

The HLT Agency is situated at the Dutch Language Union in The Hague (Netherlands), and they have an office at the University of Antwerp in Belgium.
Visit www.tst-centrale.org or www.hltagency.org for more information.

The TST-Centrale team consists of, from left to right, Michel Boekestein (technical matters), Griet Depoorter (content and service desk) and Remco van Veenendaal (project leader). Not on the photo is Linda Stokman (workflows and communication).

[i] Common Language Resources and Technology Infrastructure (www.clarin.eu)

Funded by the Department of Arts and Culture

arts and culture
Department:
Arts and Culture
REPUBLIC OF SOUTH AFRICA

Managed by the Centre for Text Technology (CTexT®) of the North-West University (Potchefstroom Campus)

NORTH-WEST UNIVERSITY
YUNIBESITI YA BOKONE-BOPHIRIMA
NOORDWES-UNIVERSITEIT
POTCHEFSTROOM CAMPUS

European Language Resource Association

At the recent official launch of the RMA, CTexT was fortunate to welcome Dr Khalid Choukri (Secretary General) of the European Language Resources Association (ELRA) as international guest speaker.

“ELRA is active in the identification, distribution, collection, validation, standardisation, improvement, in promoting the production of language resources, in supporting the infrastructure to perform evaluation campaigns and in developing a scientific field of language resources and evaluation” (www.elra.info/). They are therefore a very apt role model for the young RMA.

Dr Choukri’s speech at the launch event was informative, without excluding members of the audience who were less familiar with the concept of language technology.

He spoke of the undisputed importance of language technologies as essential facilitators to access cyberspace (for example through the Internet). Sectors of activities dependent on these resources include e-health, e-government, e-markets and e-business. This is true not only in Europe, but also globally. He also referred to applications and services that are based on language technologies, such as dictation and audio transcription, subtitling, speech-to-speech translation, car navigation and transcriptions in meetings.

“There are over 400 000 translators in the world at present, of which 150 000 are based in Europe. There is a need to translate 506 language pairs in the European Union, 110 in South Africa, and 462 in India... 6 000 languages in all. The need for translation grows by 30% every year. Automation is therefore essential,” says Dr Choukri. This cannot be done without the necessary language resources.

Dr Choukri identified the following potential cooperation and support-services for a fruitful cooperation with the RMA:

- The ELRA Catalogue and META-SHARE;
- The identification of existing resources and gaps;
- The production of new language resources (including customisation and repurposing);
- The evaluation of technologies;
- The organisation of joint workshops between the European Union and South Africa for the purposes of institutional cooperation; and
- Joint efforts for the standardisation of language resources, evaluation and best practices.

“I am very honoured to be here today to celebrate the birth of a sister organisation. We have waited (in the northern hemisphere) for so long to see this event taking place that I could not imagine missing it,” he said.

The South African Language Resource Management Agency (RMA) was founded as part of a systematic development plan to support the development of human language technologies in a multilingual environment. It is sponsored by the Department of Arts and Culture and is run by the Centre for Text Technology (CTexT®) at the North-West University (Potchefstroom Campus).

“META-SHARE is a new European language technology initiative for language resources. It comprises a distributed network of repositories and data centres. It is in essence a marketplace where language data and tools are documented, uploaded and stored in repositories, catalogued and announced, downloaded, exchanged and discussed. As it brings together several organisations and initiatives, it is an important item for international cooperation.”

http://metashare.elda.org/

Dr Khalid Choukri (ELRA)

Funded by the Department of Arts and Culture

arts and culture
Department:
Arts and Culture
REPUBLIC OF SOUTH AFRICA

Managed by the Centre for Text Technology (CTexT®) of the North-West University (Potchefstroom Campus)

NORTH-WEST UNIVERSITY
YUNIBESITI YA BOKONE-BOPHIRIMA
NOORDWES-UNIVERSITEIT
POTCHEFSTROOM CAMPUS ®