Reykjavík (Iceland), 26th May 2014

Poster Session

13:00-15:00

Chairpersons: Laurette Pretorius and Claudia Soria

An Update and Extension of the META-NET Study "Europe's Languages in the Digital Age"

Georg Rehm, Hans Uszkoreit, Ido Dagan, Vartkes Goetcherian, Mehmet Ugur Dogan, Coskun Mermer, Tamás Varadi, Sabine Kirchmeier-Andersen, Gerhard Stickel, Meirion Prys Jones, Stefan Oeter, Sigve Gramstad

This paper extends and updates the cross-language comparison of LT support for 30 European languages as published in the META-NET Language White Paper Series. The updated comparison confirms the original results and paints an alarming picture: it demonstrates that there are even more dramatic differences in LT support between the European languages.

Hungarian-Somali-English Online Dictionary and Taxonomy

István Endrédy

Background. The number of Somalis coming to Europe has increased substantially in recent years. Most of them do not speak any foreign language, only Somali, but a few of them speak English as well. Aims. A simple and useful online dictionary would help Somalis in everyday life. It should be online (with easy access from anywhere) and it has to handle billions of word forms, as Hungarian is heavily agglutinative. It should handle typos as the users are not advanced speakers of the foreign languages of the dictionary. It should pronounce words, as these languages have different phonetic sets. It should be fast with good precision because users do not like to wait. And last but not least, it should support an overview of the vocabulary of a given topic. Method. A vocabulary (2000 entries) and a taxonomy (200 nodes) was created by a team (an editor and a native Somali speaker) in an Excel table. This content was converted into a relational database (mysql), and it got an online user interface based on php and jqueryui. Stemmer and text-to-speech modules were

included and implemented as a web service. Typos were handled with query extension. Results. Although the dictionary lookup process does stemming with a web service and makes a query extension process, it is very fast (100-300ms per query). It can pronounce every Hungarian word and expression owing to the text-to-speech web service. Conclusion. This dictionary was opened to the public in October, 2013. (http://qaamuus.rmk.hu/en) The next step is the creation of a user interface optimised for mobile devices.

Computerization of African Languages-French Dictionaries

Chantal Enguehard, Mathieu Mangeot

This paper relates work done during the DiLAF project. It consists in converting 5 bilingual African language-French dictionaries originally in Word format into XML following the LMF model. The languages processed are Bambara, Hausa, Kanuri, Tamajaq and Songhai-zarma, still considered as under-resourced languages concerning Natural Language Processing tools. Once converted, the dictionaries are available online on the Jibiki platform for lookup and modification. The DiLAF project is first presented. A description of each dictionary follows. Then, the conversion methodology from .doc format to XML files is presented. A specific point on the usage of Unicode follows. Then, each step of the conversion into XML and LMF is detailed. The last part presents the Jibiki lexical resources management platform used for the project.

Morphological Analysis for Less-Resourced Languages: Maximum Affix Overlap Applied to Zulu

Uwe Quasthoff, Sonja Bosch, Dirk Goldhahn

The paper describes a collaboration approach in progress for morphological analysis of less-resourced languages. The approach is based on firstly, a language-independent machine learning algorithm, Maximum Affix Overlap, that generates candidates for morphological decompositions from an initial set of language-specific training data; and secondly, language-dependent post-processing using language specific patterns. In this paper, the Maximum Affix Overlap algorithm is applied to Zulu, a morphologically complex Bantu language. It can be assumed that the algorithm will work for other Bantu languages and possibly other language families as well. With limited training data and a ranking adapted to the language family, the effort for manual verification can be strongly reduced. The machine generated list is manually verified by humans via a web frontend.

InterlinguaPlus Machine Translation Approach for Under-Resourced Languages: Ekegusii & Swahili

Edward O. Ombui, Peter W. Wagacha, Wanjiku Ng'ang'a

This paper elucidates the InterlinguaPlus design and its application in bi-directional text translations between Ekegusii and Kiswahili languages unlike the traditional translation pairs, one-by-one. Therefore, any of the languages can be the source or target language. The first section is an overview of the project, which is followed by a brief review of Machine Translation. The next

section discusses the implementation of the system using Carabao's open machine translation framework and the results obtained. So far, the translation results have been plausible particularly for the resource-scarce local languages and clearly affirm morphological similarities inherent in Bantu languages.

UNLarium: a Crowd-Sourcing Environment for Multilingual Resources

Ronaldo Martins

We present the UNLarium, a web-based integrated development environment for creating, editing, validating, storing, normalising and exchanging language resources for multilingual natural language processing. Conceived for the UNL Lexical Framework, the UNLarium provides semantic accessibility to language constrained data, as it interconnects lexical units from several different languages, through taxonomic and non-taxonomic relations, representing not only necessary but also typical associations, obtained from machine learning and human input, in order to create an incremental and dynamic map of the human knowledge.

Collaborative Language Documentation: the Construction of the Huastec Corpus

Anuschka van ´t Hooft, José Luis González Compeán

In this paper, we describe the design and functioning of a web-based platform called Nenek, which aims to be an on-going language documentation project for the Huastec language. In Nenek, speakers, linguistic associations, government instances and researchers work together to construct a centralized repository of materials about the Huastec language. Nenek not only organizes different types of contents in repositories, it also uses this information to create online tools such as a searchable database with documents on Huastec language and culture, E-dictionaries and spell checkers. Nenek is also a monolingual social network in which users discuss contents on the platform. Until now, the speakers have created a monolingual E-dictionary and we have initiated an on-going process of the construction of a repository of written texts in the Huastec language. In this context, we have been able to localize and digitally archive documents in other formats (audios, videos, images), yet the retrieval, creation, storage, and documentation of this type of materials is still in a preliminary phase. In this presentation, we want to present the general methodology of the project.

Open-Source Infrastructures for Collaborative Work on Under-Resourced Languages

Sjur Moshagen, Jack Rueter, Tommi Pirinen, Trond Trosterud, Francis M. Tyers

In order to support crowd sourcing for a language, certain social and technical prerequisites must be met. Both the size of the community and the level of technical support available are important factors. Many language communities are too small to be able to support a crowd-sourcing approach to building language-technology resources, while others have a large enough community but require a platform that relieves the need to develop all the technical and computational-linguistic know how needed to actually run a project successfully. This article covers the languages

being worked on in the Giellatekno/Divvun and Apertium infrastructures. Giellatekno is a language-technology research group, Divvun is a product development group and both work primarily on the Sámi languages. Apertium is a free/open-source project primarily working on machine translation. We use Wikipedia as an indicator to divide the set of languages that we work on into two groups: those that can support traditional crowdsourcing, and those that do not. We find that the languages being worked on in the Giellatekno/Divvun infrastructure largely fall into the latter group, while the languages in the Apertium infrastructure fall mostly into the former group. Regardless of the ability of a language community to support traditional crowdsourcing, there is in all cases the necessity to provide a technical infrastructure to back up any linguistic work. We present two infrastructures, the Giellatekno/Divvun infrastructure and the Apertium infrastructure and show that while both groups of language communities would not be able to develop language technology on their own, using the infrastructures that we present they have been quite successful.