Thoughts on Multilingual Language Resources for Under-resourced Languages as Linked Data

Laurette Pretorius

2nd LIDER Roadmapping Workshop
Madrid
9 May, 2014

UNISA | college of graduate studies

Structure of the presentation

- LTs and LD: Interest
- Bottlenecks and Challenges in the way we currently publish, share and Consume LRs?
- Potential of LD in improving the way in which we publish, share and consume LRs?
- Specific Issues?

LTs and LD: Interest

- **LTs**: The technological development of the indigenous South African languages, including Zulu, Xhosa, Swati, Ndebele, Tswana, and Afrikaans.

- **LD**: Using LD in weaving the under-resourced languages of Southern Africa into the fabric of the Multilingual Semantic Web

- **Under-resourced languages**: Languages that have a small or economically disadvantaged user base, that are therefore typically ignored by the commercial world and that are technologically under-developed due to limited human, financial and linguistic/language resources.

- **Specific interests**: Exposing existing material as LD (multilingual terminologies, translation memories, linguistic data and multilingual parallel corpora for various Southern African languages, indigenous knowledge in indigenous languages).

Bottlenecks and Challenges in the way we currently publish, share and consume LRs?

Towards multilingual LRs and applications that would facilitate cross-lingual information production and consumption in the Semantic Web, ***also for diverse, under-resourced languages***:

- **Publishing LRs**: Development, standardisation and quality assurance of such LRs.
- **Sharing LRs** (both people sharing their LRs or closely-related languages sharing LRs): Availability, accessibility, interoperability and optimal exploitation of cross-linguistic and cross-language similarities of LRs.
- **Consuming LRs**: Identification and access of suitable and appropriate LRs across languages and domains.

Potential of LD in improving the way in which we publish, share and consume LRs?

*In a philosophical sense LD allows the whole to be greater than the sum of its parts, also for LRs for **under-resourced languages**:*

- **Finding LRs**: Increasingly sophisticated search capabilities over existing LRs that are published as LD.
- **Publishing LRs**: Exposing data in under-resourced languages in the Semantic Web through LD without the obligation to have extensive language technology support in place.
- **Sharing LRs**: LD to allow for the flexible, scalable sharing of LRs, and thereby offers new opportunities for language technologies across languages.
- **Consuming LRs**: By providing multiple interlinked contexts, and linguistic and language diversity, LD has the possibility to profoundly impact on the nature and quality of language technologies that use these LRs.

Specific Issues?

- At a very practical level, a first bottleneck in exposing data, available in new languages, is often identifying the right URI for the concept/term at hand - semi-automated approaches would be most useful – going from "four star status" to "five star status"

- Cross-lingual production and consumption of, and reasoning over Small Data in new, often under-resourced languages of choice, for specific sectors and/or domains of choice.

- A word of thanks and expectation: We look forward to the outcomes of projects such as LIDER in terms of their significance for real language diversity in the Semantic Web.