COLLABORATING AROUND LANGUAGE RESOURCES FOR THE DIGITAL SINGLE MARKET

An LT_Observatory Workshop
co-organised with LT-Innovate & ELRA/ELDA

Date: 8 December 2016 - **Venue**: ELRA/ELDA - 9, Rue des Cordelières - 75013 Paris

BRIEFING

Overall purpose of this workshop

- Bring together a core group of stakeholders in **LRs** (language resources – i.e. digital parallel resources used for translation - see below**) identified during the LT_O project **to create a plausible agenda of the next three to five years for building intelligent LR management** (also see below*) **into the "MT infrastructure" in the EU**.
- Use this agenda to explore/define **responsibilities, opportunities, collaboration structures, and research-to-innovation** projects with respect to LR futures.

Background assumptions

Data-driven translation technology is now reaching a critical mass of awareness and performance. It uses some form of "deep" machine learning that operates on textual (or spoken) data to discover significant patterns of cross-language equivalence.

Other emerging language knowledge technologies (word, phrase and conversational semantics, text analytics, etc.) will hopefully extend the effectiveness and scope of this baseline MT technology. There is a considerable body of work in Europe ongoing in this field. But also in SE Asia, India, and the Americas, so competition will be tough.

The end result should be ubiquitous MT deployment across Europe, drawing on language resources, technology expertise and marketing through LSPs, govt. services, online services, commercial platforms, and mobile MT services.

This MT as a ubiquitous service will depend on seamless access to the right LR fuel supply to facilitate engine building.

Financial support must be found for this MT research & innovation dynamic. If EC, govt. or venture capital money is able to fund innovation efforts on top of the research in order to industrialise, simplify, streamline, lower the entry cost and generally disrupt the existing paradigm of fragmented, complicated, much-contested MT use, then we as LR stakeholders will be able to open up new

opportunities for MT deployment. This is especially true for "language-resourcing" the digital single market in Europe.

Additional *new* technologies that might emerge in the R&I galaxy are:

- Self-learning MT systems, ("cognitive MT"?)
- Real-time translation using ready-made engines (previously constructed using existing LRs)
- Spoken-language MT systems with added *affective computing* (understanding and representation of emotional dimensions of human discourse across different languages),
- Automated speech-to-text captioning of videos etc.
- Dedicated "vertical" MT systems focused on domains/industries, use cases and text types.
- LR search engines (virtual assistants) that could automatically search for (and eventually create) appropriate LRs (e.g. a search for LRs that include lots of interrogative forms, or vehicle parts lists, or names of diseases or other definable categories for specific translation tasks).

(***"LR management"** here means: identifying, evaluating, improving or creating, and making easily available appropriate LRs for end-user needs across all relevant EU language pairs at lowest cost and maximum quality. Ideally by automating (some of) the supply pipeline)

(****LRs** = parallel language corpora, comparable language corpora, terminology and related knowledge bases, and monolingual language corpora in textual form. This criterion could eventually be extended to other modalities such as spoken language, and include support for language understanding using computer understanding of images, sounds and video sequences in certain situations).

Key issues & questions to be covered in the workshop

1. **Building effective relations between the EC's CEF.AT programme and the commercial MT user community for access to existing and new LRs**

   - What if LR access/quality became the criterion for choosing an MT supplier?
   - Who has carried out a convincing needs analysis of the community of MT users? If so how are needs for LRs expressed?
   - Should the community be trying to "educate" end users as to the validity, benefit and operational feasibility of using MT?
   - Will ELRC (European Language Resource Coordination) be able to share discovered or new LRs with the commercial community in general? If so under which legal/financial/technical conditions?
   - Will ELRC be able to apply LT_0's criteria for "operational usability"*** in LRs?
   - Does the EU research community have sufficient vision, resources and momentum to address the linguistic challenge of building an LR hub for **24 to 50 EU languages within 5 years** to build on this MT R&D momentum at least for the EU marketplace? Can the CEF and HORIZON 2020 programmes help deliver on this promise?

2. **Expanding and accelerating LR discovery/creation in Europe**

   - How can **Open Data in Europe** be harnessed to aid the production of new, relevant LRs?
   - Which **research focus** will be most needed to boost innovation in tomorrow's LR management?
   - What kind of **pipeline** can best streamline the innovation step? Classic EC projects, or some new vehicle? Could there be a LR marketplace focused on MT requirements?

- What measures can be introduced to **accelerate LR-MT uptake** in government and business in partnership with R&I efforts?
- What role can **LSPs** (language service suppliers –translation companies) play in this value chain?

(*** **Operational Usability** of Language Resources means being able to freely access:
- Links from LT Observe to parallel language resources that match an end-user's needs
- Positively evaluated by experts
- Featuring appropriate alignment (segments or sentences or paragraphs)
- Containing domain-related text
- Or domain-relevant vocabulary/terminology lists
- Either free of charge or at a reasonable price)
- Bearing valid metadata, including production date, ownership and contact information
- All listed in the LT Observe repository at http://www.lt-innovate.org/lt-observe/resources-list.)