13.

Sue Ellen Wright, Menzo Windhouwer, Ineke Schuurman, Marc Kemps-Snijders

Foreword

It's a pleasure to write the Foreword for the book on Collaboratively Constructed Language Resources.
I believe that the trend of collaborative construction of Language Resources (LRs) represents both a "natural" evolution of computerised resource building (I'll try to give few historical hints) and a "critical" evolution for the future of the field of language resources.

Some historical hints

Where does collaborative resource construction position itself in the language resource field?
I'll just give a glimpse here at some historical antecedents of the current collaborative methodology, without mentioning the obvious ones, like Wikipedia or Wiktionary.

Where does collaborative resource construction position itself in the language resource field?
I'll just give a glimpse here at some historical antecedents of the current collaborative methodology, without mentioning the obvious ones, like Wikipedia or Wiktionary.

19th century lexicographic enterprise

We have not invented collaborative construction of language resources, or even crowdsourcing, just recently.

George P. Marsh used it already in 1859 for the Philological Society of London for "the preparation of a complete lexicon or thesaurus of the English language", the New English Dictionary (now known as the Oxford English Dictionary).
Acting as Secretary in America he decided to "adopt this method of bringing the subject to the notice of persons in this country who may be disposed to contribute to the accomplishment of the object, by reading English books and noting words …". Moreover: " … the labors of the English contributors are wholly gratuitous".

Given that not much material was collected after this appeal, a similar appeal (http://public.oed.com/history-of-the-oed/archived-documents/april-1879-appeal/) was re-launched twenty years later, in 1879,
by the dictionary's editor James Murray when "volunteer readers were recruited to contribute words and illustrative quotations":
"… the Committee want help from readers in Great Britain, America, and the British Colonies, to finish the volunteer work so enthusiastically commenced twenty years ago …",
and "A thousand readers are wanted, and confidently asked for, to complete the work as far as possible within the next three years, so that the preparation of the Dictionary may proceed upon full and complete materials."

We can't deny that this is a clear example of collaborative construction of a language resource! It could even be defined as an early example of crowdsourcing.

More recent examples: some EC resource projects of the 20th century

Other – more recent – examples could be found in the policy adopted in projects funded by the European Commission, in the '90s, where many language resources had to be collaboratively built inside a consortium of many partners. Also because of this "enforced" collaboration, some features and trends presenting clear connections with the current notion of "collaborative building" emerged in the first half of the '90s:

1. The need to build a core set of LRs, designed in a harmonised way, for all the EU languages
2. The need to base LR building on commonly accepted standards
3. The need to make the LRs that are created available to the community by large, i.e. the need for a distribution policy (at that time we introduced the notion of distributing resources, not yet sharing them!).

By the way, these requirements are strictly implied by and related to the emerging notion in the '90s of the "infrastructural role" of LRs.

I just mention two types of collaborative resource building in EC projects, representing two partially different building models.

One method could be represented by the EuroWordNet projects: each partner was building the WordNet for her/his language, all modelled on – and linked to – the original Princeton WordNet, and altogether constituting a homogeneous and interrelated set of lexicons.

Another method is represented by projects like PAROLE and SIMPLE, for the construction and acquisition of harmonised resources. They were, to my knowledge, the first attempt at developing together medium-size coverage lexicons for so many languages (12 European languages), with a harmonised common model, and with encoding of structured semantic types and syntactic (subcategorisation) and semantic frames on a large scale. Reaching a common agreed model grounded on sound theoretical approaches within a very large consortium, and for so many languages, was in itself a challenging task. The availability of these uniformly structured lexical and textual resources, based on agreed models and standards, in so many EU languages, offered the benefits of a standardised base, creating an infrastructure of harmonised LRs throughout Europe.

What was interesting was that these projects positioned themselves inside the strategic policy – supported by the EC – aiming at providing a core set of language resources for the EU languages based on the principle of "subsidiarity". According to the subsidiarity concept, the process started at the EU level continued at the national level, extending to real-size the core sets of resources in the framework of a number of National Projects.

This achievement was of major importance in a multilingual space like Europe, where all the difficulties connected with the task of LR building are multiplied by the language factor. All the various language resource projects determined also the beginning of the interest in standardisation in Europe. It was seen as a waste of money, effort and time the fact that every new project was redoing from scratch the same type of (fragments of) LRs, without reusing what was already available, while LRs produced by the projects were usually forgotten and left unused. From here, the notion of "reusability" arose. As a remedy, a clear demand for interoperability standards and for common terms of reference emerged.

same type of (fragments of) LRs, without reusing what was already available, while LRs produced by the projects were usually forgotten and left unused. From here, the notion of "reusability" arose. As a remedy, a clear demand for interoperability standards and for common terms of reference emerged.

Reusability and integration of language resources

Other requirements with connections to collaborative construction of LRs are the possibility to reuse and integrate different language resources.

LRs (i.e. data) started to be understood as critical to make steps forward in NLP already in the '80s, marking a sort of revolution with respect to times and approaches in which they were even despised as an uninteresting burden. The 1986 Grosseto (Tuscany) Workshop "On automating the lexicon" was the event marking this inversion of tendency and the starting point of the process which gradually brought the major actors of the NLP sector to pay more and more attention to so-called "reusable" language resources.

In 1998, in a keynote talk at the 1st LREC in Granada, I could state that "Integration of different types of LRs, approaches, techniques and tools must be enforced" as a compelling requirement for our field: "The integration aspect is becoming – fortunately – a key aspect for the field to grow. This is in fact a sign of maturity: today various types of data, techniques, and components are available and waiting to be integrated with not too great an effort. I believe that this integration task is an essential step towards ameliorating the situation, both in view of new applicative goals and also in view of new research dimensions. The integration of many existing components gives in fact more than the sum of the parts, because their combination adds a different quality."

Among the combinations to be explored I mentioned: interaction between lexicon and corpus, integration of different types of lexicons, of various components in a chain (what we call today workflows), of Written and Spoken LRs towards multimedia and multimodal LRs, and also integration of symbolic and statistical approaches. I observed that "a single group simply does not have the means, or the interest, to carry them out. … everything is tied together, which makes our overall task so interesting – and difficult. What we must have is the ability to combine the overall view with its decomposition into manageable pieces. No one perspective – the global and the sectorial – is really fruitful if taken in isolation. A strategic and visionary policy has to be debated, designed and adopted for the next few years, if we hope to be successful."

Collaborative construction of LRs is linked to and is an evolution of both the reusability notion and the integration requirement.

Language Technology as a data intensive field: the data-driven approach

LRs were not conceived as an end in themselves, but as an essential component to develop robust systems and applications. They were the obvious prerequisite and the critical factor in the emergence and the consolidation of the data-driven approach in human language technology. Today we recognise that Language Technology is a data-intensive field and that major breakthroughs have stemmed from a better use of more and more Language Resources.

From Murray appeal, through Corpus-based lexicography, back to collaborative work!

In the '90s computer-aided corpus-based lexicography became the "normal" lexicographic practice for the identification and selection of documentation – through text-processing methods, frequency lists, patterns spotting, context analysis, and so on. No need to ask for 10,000 contributors!

Data-driven methods and automatic acquisition of linguistic information started in the late '80s with the ACQUILEX project, aiming at acquiring lexical information from so-called machine-readable dictionaries. The needs of "language industry" applications compelled to rely on actual usage of languages, as attested in large corpora, for acquiring linguistic information, instead of relying on human introspection as the source of linguistic information and of testing linguistic hypotheses with small amounts of data. This meant developing statistical techniques, machine learning, text mining, and so on.

All this was/is very successful, but all these techniques on one side rely on bigger and bigger collections of data (LRs), possibly annotated in many ways and often with human intervention, and on the other side they are never 100% correct, thus requiring again human intervention. Therefore, if more and bigger (processed) LRs are needed, if statistical techniques arrive at a certain limit, new ways to cope with this need of "Big Data" must be found and explored. Natural ways of coping with the big data paradigm and the need of accumulation of extremely large (linguistic) knowledge bases are:

1. collaborative building of resources on one side, and
2. putting again human intelligence in the loop on the other side, recognising that some tasks are better performed by humans: crowdsourcing as a form of global human-based computation.

Collaborative building vs. crowdsourcing can be paralleled to the difference between involvement and contribution of colleagues (as in the EC projects above) vs. involvement of the layman/everyone (as in Murray appeal). Even if both can be said to rely on collective intelligence or on the "wisdom of the crowd", they clearly represent quite different approaches and methodologies and require different organisations

Collaborative building vs. crowdsourcing can be paralleled to the difference between involvement and contribution of colleagues (as in the EC projects above) vs. involvement of the layman/everyone (as in Murray appeal). Even if both can be said to rely on collective intelligence or on the "wisdom of the crowd", they clearly represent quite different approaches and methodologies and require different organisations.

Language Resources and the Collaborative framework: to achieve the status of a mature science

The traditional LR production process is too costly. A new paradigm is pushing towards open, distributed language infrastructures based on sharing LRs, services and tools. Joining forces and working together on big experiments that collect thousands of researchers is – since many years – my dream, what I think is the only way for our field to achieve the status of a mature science.

It is urgent to create a framework enabling effective cooperation of many groups on common tasks, adopting the paradigm of accumulation of knowledge so successful in more mature disciplines, such as biology, astronomy, physics. This requires enabling the development of web-based environments for collaborative annotation and enhancement of LRs, but also the design of a new generation of multilingual LRs, based on open content interoperability standards. The rationale behind the need of open LR repositories is that accumulation of massive amounts of (high-quality) multi-dimensional data about many languages is the key to foster advancement in our knowledge about language and its mechanisms. We must finally be coherent and take concrete actions leading to the coordinated gathering – in a shared effort – of as many (processed/annotated) language data as we are collectively able to produce. This initiative compares to the astronomers/ astrophysics' accumulation of huge amounts of observation data for a better understanding of the universe.

Consistently with the vision of an open distributed space of sharable knowledge available on the web for processing, the "multilingual Semantic Web" may help in determining the shape of the LRs of the future and may be crucial to the success of an infrastructure – critically based on interoperability – aimed at enabling/improving sharing and collaborative building of LRs for a better accessibility to multilingual content. This will serve better the needs of language applications, enabling building on each other achievements, integrating results, and having them accessible to various systems, thus coping with the need of more and more 'knowledge intensive' large-size LRs for effective multilingual content processing. This is the only way to make a giant leap forward.

Relations with other dimensions relevant to the LR field

In the "FLaReNet Final Blueprint", the actions recommended for a strategy for the future of the LR field are organised around nine dimensions: a) Infrastructure, b) Documentation, c) Development, d) Interoperability, e) Coverage, Quality and Adequacy, f) Availability, Sharing and Distribution, g) Sustainability, h) Recognition, i) International Cooperation. Taken together, as a coherent system, these directions contribute to a sustainable LR ecosystem.

Let's not forget that the same requirements apply whatever the method of LR building: collaboratively built resources undergo the same rules/recommendations. An implication of collaboration is that interoperability acquires even more value. The same is true for sustainability, for data infrastructure enabling international collaboration, and also for notions such as authority and trust. Moreover, when collaborative building is explicitly performed, there is the need to better define all the small steps inside an overall methodology. These recommendations could be taken as a framework in which to insert our future work strategy also in the collaborative paradigm.

Let's organise our future!

One of the challenges for the collaborative model to succeed will be to ensure that the community is engaged at large! This can also be seen as an effort to push towards a culture of "service to the community" where everyone has to contribute. This "cultural change" is not a minor issue. This requirement was for example at the basis of the LRE Map idea, a collaborative bottom-up means of collecting metadata on LRs from conference authors, contributing to the promotion of a large movement towards an accurate and massive bottom-up documentation of LRs (see also http://www.resourcebook.eu/ with metadata for about 4000 LRs from many conferences)

My final remark is that, as with any new development, it is important on one side to leave space to the free rise of new ideas and methods inside the collaborative paradigm, but is also important to start organising its future. There must be a bold vision and an international group able to push for it (with both researchers and policy makers involved) and to organise some grand challenge that, via a distribution of efforts and exploiting the sharing trend, involves the collaboration of a consistent portion of our community. Could we envision a large "Language Library" as the beginning of a big Genome project for languages, where the community collectively deposits/creates increasingly rich and multi-layered LRs, enabling a deeper understanding of the complex relations between different annotation layers/language phenomena?

Pisa, Italy Nicoletta Calzolari