Very large language resources? At our finger!

Dan Cristea
"Alexandru Ioan Cuza" University of Iasi
16, Berthelot St., 700483 – Iasi, Romania
E-mail: dcristea@info.uaic.ro

Abstract

The paper proposes a legislative initiative for acquiring large scale language resources. It militates for raising a large awareness campaign that would allow the storing and preservation for research purpose, in electronic form, of all textual documents which go to print in a country.

1. Introduction

This paper brings into attention a proposal for conserving over long time and using largely, at a national level, for research purposes, the linguistic data which are printed and distributed for public use daily by editorial houses.

It is evident that, without a continuous effort, those languages which are now called "less-resourced" will continue to be viewed like that even when, hypothetically, they will promote to the same amount of resources as the languages that at this very moment are known to be most resourced. Moreover, if the most resourced languages would cease to acquire resources now, on the ground that they have fulfilled their needs, in short time they will lose their leading positions. This is because LRs become obsolete very quickly. Even more, if we look at the annotated resources, the linguistic facts which are subject to automatic annotation could change over time, as the linguistic theories on which the marking conventions are based evolve, and as the automatic annotation processes themselves get improved. So, as the language goes along and evolves and our vision with respect to the language changes, the resources, themselves, get old. There is no end in building LRs.

In many countries a "legal deposit" law is in use. It obliges all providers of printing materials (editing houses, physical or juridical persons which print documents for public, recording houses and studios, the National Bank, the State Mint, the National Post, etc.) – let's call them *resourcers* – to send a number of copies of each printed item intended for distribution to a national library (which could be one physical unit or a consortium of libraries) for long-time preservation. Although the horizon of media production changed dramatically in the last years, to my knowledge, there are only very timid trials for improvement of the juridical aspects.

As resources are needed dramatically and many of them are very expensive, the issue of acquiring them should stop from being accidental or episodic and should become a national policy. Something should be done. A law should defend the linguistic resources of the languages spoken in a country as being of primary interest. This paper discusses one possible solution which, although not simple to implement, could change completely the LRs scene in the near future.

2. Enhancing the legislation on legal deposits

A recent investigation among some of the most important producers of printed information in Romania revealed that many editing houses are keen to donate their resources for research purposes. However, another fraction, which unfortunately makes the majority, is not interested to collaborate. They ignore the importance of the issue, are fearful that donating their data is equivalent to loosing the property control over them, will possibly trigger a loss of profit, or simply do not have time to dedicate to this kind of matters.

In reality, nothing of the kind has to happen. Although we need their linguistic data, we do not want the *resourcers* to be harmed if they give their data to science. The idea is to promote a legislative initiative that imposes the compulsoriness for the *resourcers* to donate their linguistic data for language research. The proper moment has come to try to raise the awareness for a concentrated action in Europe. We need to raise governmental interest towards the promotion of such legislation, simultaneously in many countries.

The following type of resources, produced in series, would be in focus to such a law, irrespective whether the resources are intended for commercial or for free distribution: books, booklets, leaflets, journals, magazines, almanacs, calendars, musical scores, propagandistic materials having a political, administrative, cultural, artistic, scientific, educational, religious, a.s.o. goal, posters, proclamations, any other materials intended for publication on public places, Ph.D. thesis, university courses, documents in electronic format containing linguistic material (CDs, DVDs, etc.), standards and technical norms, publications issued by national and local authorities, collections of norms and laws, any other printed or multiplied material by using graphical or physical-chemical methods.

On the practical level, the initiative presupposes the existence of a national repository, which is an entity (IT center, institute, etc. – let's call it the *Portal*), which, on one hand, has the legal authority to receive and store data contributed by *resourcers*, and, on the other hand, is technically equipped to collect and record, indefinitely

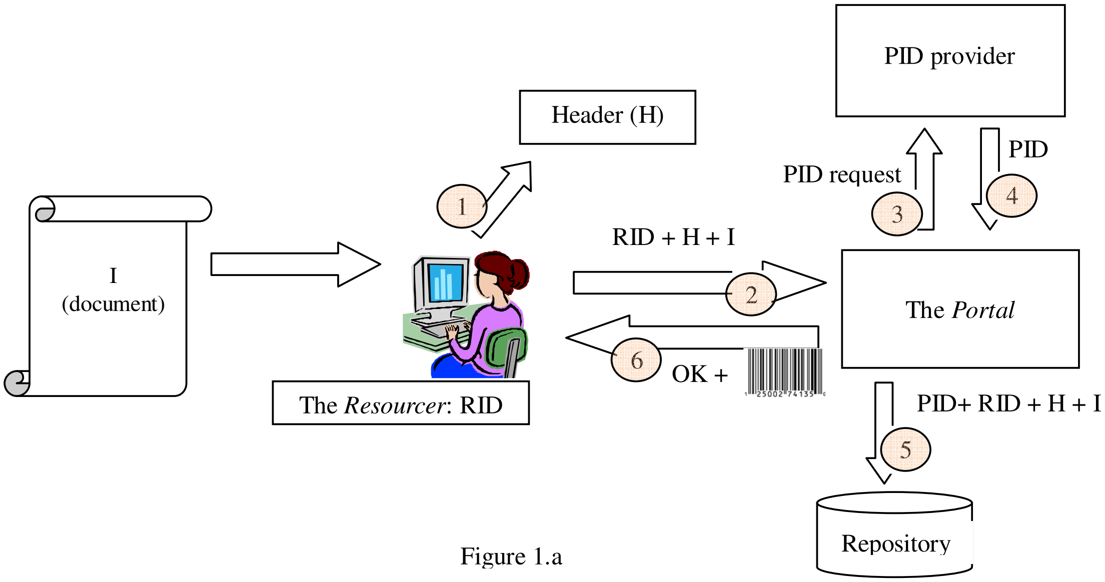

Figure 1.a

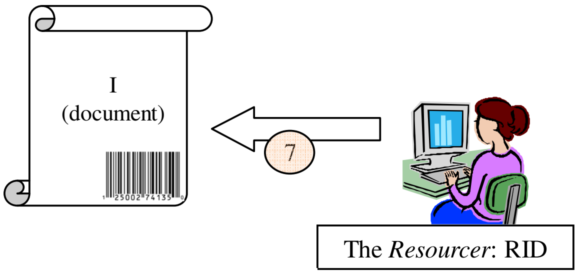

Figure 1.b

long, in electronic format, all data issued for publication, daily, in a country.

The law should state that by sending an electronic copy for long-time preservation to this national repository no authoring rights or commercial benefits are lost by the *Resourcer*. The copy can be used, intermediated by the *Portal*, only for research purposes applied to language and the *Portal* cannot make public the data on internet or on other media, unless it is asked to do so by the owner. It is clear that a fragile IPR chapter will not be acceptable in the text of this law (COM, 2009). A weak statement of IT security measures to protect the authors' rights will also be amendable. All these aspects are very important and should receive full attention in the formulation of this law.

3. The capturing flow

I see the *Portal* as a factory that processes words. The start elements of the data flow should be as follow: before issuing the first publication, or at the moment the law is imposed, the *Resourcer* should have got an identification code (*RID*) from the *Portal*. It will use this code for communication with the *Portal* regarding any publication, during all its juridical lifetime.

Suppose that today the *Resourcer* prepares for publication a new item *I*, which has got the "ready for printing" editorial approval. The Figure 1.a explains the communication initiated by this new item. The *Resourcer* fills in an electronic form (header – *H*), containing identification information of the document, and then interacts with the *Portal*, uploading its *RID*, the header *H* and an editable copy of *I*. The *Portal* receives this data and asks for a persistent identification code (*PID*) to an authority capable of issuing them (Kunze and Rogers, 2003; Schwardmann, 2009). When it gets one, it stores in its repository a bunch of data containing *PID*, *RID*, *H*, and *I*. Then, the *Portal* returns to the *Resourcer* an OK message, containing two parts: a human readable part and a bar code part. The OK box should record a seal of the *Portal*, together with the *PID* and the *RID*. Now the document, which has also this OK box, included on an inner cover or on a sleeve, can be printed (Figure 1.b). This box proves to any authority in charge of controlling the application of the law that the legal deposit was performed by the *Resourcer* on the *Portal*, and all the needed identification information is there.

The above detailed exchange of data between the *Resourcer* and the *Portal*, including also a communication with a third entity responsible for issuing PIDs, seems heavy and time consuming, and if so, totally unacceptable by the editing houses. Indeed, it is a known fact that these entities are most of the time constrained to process data at great speed, especially, if they print daily newspapers, for example. Nevertheless, the communication which, as is described above, appears to be heavy and cumbersome, can be done as quickly as a blink of an eye by making completely automatic the whole chain, including the fill-in of identification information contained in the header *H*. The content of the header can be extracted by specialized modules from the electronic item *I*. So, practically, the entire chain could be activated by a click on a button on the editing interface. This should end up, almost instantly, with the inclusion of the OK box in a dedicated place of the document going to print.

4. Data processing

Once captured, data on the *Portal* should be processed. In this section I describe a list of processing capabilities that the *Portal* should be able to provide.

First, it is obvious that the Portal should have sufficient storing capacities and that these capacities should be specially designed for preserving data indefinitely long periods of time. Then if should display indexing, search and retrieval capacities, at different levels: header, lexical tokens (words), lexical expressions, as well as contextual information. This means that each document, once placed on the portal, should be submitted to a processing chain that includes, minimally: tokenization, part-of-speech tagging, lemmatization and indexing. It is foreseeable therefore that each document will be recorded as raw text on which the standoff XML annotation will make reference. The XML annotation and the indexing requirements will most probably multiply the size of the initial text documents a couple of times.

Based on these basic functionalities, a different line of processing refers to lexicographic needs. The Portal should be able to perform complex operations such as: detection of foreign words, signaling of new words, recognition of senses of words in context (WSD), detection of new senses, signaling of forgotten (obsolete) words, signaling of senses which are no more used, etc. For instance, signaling of new words and of forgotten (obsolete) words should be triggered by a frequency of occurrence which, over a given interval of time, is above/below certain thresholds, as decided by a linguistic authority. Similarly, signaling of a new sense could be triggered by the fail to align the sense recognized in context to those kept in a repository of senses, like for instance an authoritarian explanatory dictionary, if this happens with a certain frequency recently, and if the pattern of use is sufficiently stable. Forgotten (obsolete) senses are recognized by the occurrence of these senses under a certain threshold.

The process which should be placed at the base of recognizing obsolete words or senses presupposes placing a bag of words under constant surveillance. These are words/senses plausible of becoming under-used because they experience a constantly degrading frequency. Let's note that the criterion of absolute or even relative frequency, over a certain interval, could prove not being relevant, because there are words which are very rarely used, although they could not be in danger of being considered extinguishable (some science neologisms, for instance). The best way to do this is to associate to each word a personal file, recording a set of dynamic features, among which the frequency of occurrence over time (a graphic, from which a gradient of deterioration could be computed), the list of registers that use it (with the associated relative frequencies), etc. So, the problem resides in computing the frequency over a constant interval of time, considered always back from the current day. One could do this by simply searching the spotted word in the repository and counting only the occurrences that fall in the needed interval – a function that would be called only once in a certain long interval – say two to five years (because one cannot expect that the tagging "obsolete" can be updated too frequently, from yesterday to today…).

It is clear that any decision on anyone of these positions should ultimately be taken by a linguistic authority (Academia). Their decisions should investigate the signals transmitted by the Portal, which are rooted on neat statistical evidence.

Different processing flows could implement other functions. A number of resources, which are of increasing importance in keeping a language technologically updated, can be continuously connected onto the Portal. Among these, I see: the main Dictionary of the language, the WordNet (Fellbaum, 1998), the VerbNet (Kipper et al., 2008), the FrameNet (Fillmore, 1976; Atkins et al, 2003) – to name just a few. Supposing all these resources are complete for the language L, at a certain moment, they should be kept updated with the evolution of language. So, any dynamics in language should be mirrored in these resources as well. If, as suggested above, each lexical item of the language has a personal record on the Portal, then if should include references in all these resources. As such, the word *w* is linked to its input in the Dictionary, where the inventory of senses is recorded, and these senses are aligned to those listed in the WordNet for this lexical item, as well to its entry in VerbNet and FrameNet. All these resources are connected among them and kept online with the evolution of language by the Portal.

The Portal can host also a number of services addressed to the *resourcers*, to the language researchers, to the consumers or to the public at large. Public services could be charged to the customers and benefits be returned to the *resourcers*, in amounts proportional to their monthly contribution on the Portal (measured in characters).

Other types of paid services could be imagined, with benefits returned to the *resourcers*, for instance advertising publications and on-line access to parts of their publications, which they are keen to offer on the market. The possibility to develop a set of services from which the *resourcers* could obtain profit is interesting also from the point of view of potentially lowering the *resources*' opposition vis-a-vis of a law that would impose the obligation of continuous language preservation, as has been discussed in section 2.

5. Evaluation

It is clear that the type of processing encumbered by such an initiative would bring to the *Portal* a very big amount of linguistic data daily. A rough evaluation of the processing needs and costs encumbered by such a national-wide enterprise should bring into focus parameters such as: the number of editorial houses registered, the average number of publications of a publishing house per year, the average length in pages of a printed item, the average number of characters per page. Leaving aside episodic publications of small size, our enquiry about the average amount of data published in books and journals, in a medium size country of Europe (Romania), at the level of the year 2008, has yielded an amount of textual data which is less than 1Gb daily.

A channel with a bandwidth of 12.5 Mb/sec can lightly face the required transfer described in section 3, avoiding bottlenecks on moments of crowd. Load balancing and mirroring, for safety reasons, should be assured, by storing the data on at least two centers, in different locations. As proved already by data intensive storing houses (Google [1], for instance), software RAID technology, made up of a farm of small computers, is a cheap and appropriate solution for long time preservation and a comfortable processing speed.

6. Conclusions

The advantages of a *Portal* able to process linguistic data at a scale as the one envisioned above are hard to depict now correctly. First of all, it will give a long-time and complete solution to the problem of linguistic data preservation for the language(s) of a nation, as well as an almost complete radiography of its diachronic evolution. Secondly, it will put the basis for an exhaustive research related to language. Thirdly, it could bring into focus a large scale of commercially appealing applications, in the benefit of the authors of the texts or the *resourcers*.

The success of such an initiative at national level depends very much on a large concentrated vision. The new and very fresh breath that is being felt at this moment in Europe with respect to building language processing infrastructures, to establish standards for representation of linguistic data, and to foster large scale initiatives for the acquisition of linguistic resources, as motored by recent consortiums like CLARIN[2], FlareNet[3], T4Me[4], Meta-Net[5], etc. should also move forward a favorable legislation. The proposal advanced in this paper is also in line with other initiatives that try to raise the awareness on the necessity of free access to science[6]. It, however, does not advocate against intellectual property (Stephan, 2001), but is very much in favor of a reconsideration of the IPR legislation, which is too restrictive in many cases of usage of language resources for research. After all, our language, as we use it today, represents a collective contribution and is due to a perpetual reshaping from all its speakers from the beginning of the time… Donating his linguistic creation for language preservation and research, while not harming at all its creator, neither intellectually, nor commercially, represents just the minimum return that an author which uses the language owes to those who have invented it, for the benefit of those which will use it in the future.

7. References

Atkins, S., Rundell, M. and Sato, H. (2003) The Contribution of Framenet to Practical Lexicography, *International Journal of Lexicography*, Volume 16.3: 333-357.

COM (2009) 532 – Communication from the Commission. Copyright in the Knowledge Economy.

Fillmore, C. J. (1976): Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, Volume 280: 20-32.

Kipper,K., Korhonen, A., Ryant, N., Palmer, M. (2008) A Large-scale Classification of English Verbs, *Language Resources and Evaluation Journal*, 42(1), pp. 21-40, Springer Netherland.

Kunze, J. and R.P.C.Rogers (2003). The ARK Persistent Identifier Scheme. Internet draft at http://www.cdlib.org/inside/diglib/ark/arkspec.pdf

Schwardmann, U. (2009). PID System for eResearch. EPIC – the European Persistant Identifier Consortium, personal communication at NEERI-09, Helsinki.

Stephan, K. "Against Intellectual Property". *Journal of Libertarian Studies* 15.2 (Spring 2001): 1-53.

Fellbaum, C. (1998) WordNet: An Electronic Lexical Database, MIT Press.

[1] http://infolab.stanford.edu/~backrub/google.html
[2] www.clarin.eu
[3] http://www.flarenet.eu/
[4] http://t4me.dfki.de/
[5] http://www.meta-net.eu/

[6] See, for instance, the Washington D.C. Principles For Free Access to Science at http://www.dcprinciples.org/statement.pdf, the Open Access initiative http://www.eprints.org/openaccess/, the American Scientist Open Access Forum http://amsci-forum.amsci.org/archives/American-Scientist-Open-Access-Forum.html, The SPARC Open Access Newsletter (see an issue at http://www.earlham.edu/~peters/fos/newsletter/01-02-10.htm), the Budapest Open Access Initiative http://www.soros.org/openaccess,