The Role of Internet in Creating, Financing and Integrating Language Resources

Heiki-Jaan Kaalep
Department of General Linguistics
University of Tartu
Tartu, ESTONIA
[hkaalep@psych.ut.ee]

Rene Prillop
Department of Mathematics
University of Tartu
Tartu, ESTONIA
[prr@psych.ut.ee]

Epp Ehasalu
Department of Estonian
University of Tartu
Tartu, ESTONIA
[ehasalu@pi.estnet.ee]

Abstract

The article describes the manifold influence of Internet on creating and using language resources: language resources that are made available via Internet tend to be in a standard format; it is relatively easy to get sponsor money for creating language resources if they are intended to be available in the Internet; third parties often put material on-line which appears to be a language resource without the creators being aware of it; completely new applications, taking advantage of on-line language resources, become available (like translation aids); the freedom to use language resources in Internet is less hindered by the copyright problems. At present the possibilities of Internet in using language resources have not been fully realized. Internet is mostly used as a medium for transporting language resources from one local point to another. At the same time, Internet is an environment allowing smooth integration of language resources, e.g. various on-line dictionaries via their query forms; text documents with on-line language resources via hyperlinks; text corpora with dictionaries via hyperlinks. The existing technology allows for more applications than are available at the moment; it also allows for more sophisticated applications.

Introduction

We define language resources (LR) here as any kind of language material that helps in language technology (LT) development regardless of the original intentions of the creators of the LR. It is true that LR are often in a non-standard form, it is costly to create and convert them and finally, that there are too few of them. The last is especially true for small languages that have to take advantage of every single resource they have. The current paper describes how LT and LR developing can be promoted using Internet, referring to experience from Estonia.

Problems Inherent to LR

Once we get over the problem of having no LR at all, we immediately face two new problems:

1. LR are often incompatible with each other. E.g. dictionary encoding formats are different, program modules need different software and/or hardware platforms.
2. The creators of the LR are reluctant to give away the source text (if the LR is a dictionary) or the source code (if the LR is program module). As a rule, the authors may give a permission to use the program or lexicon as it is, but it's much harder to get a permission to use the same resource in some new tool, especially if this means uncovering of (some of) the inner structure of the LR.

How could we find a compromise between the need to have standardized LR and the desire of the developers of LR to work without constraints; between the wish to have LR freely available and the reluctance of the owners to give them away?

Fortunately there is at least one point where the developers and users of LR tend to agree: they have to be usable in Internet. As we see later, the requirements imposed on LR by Internet give a basis for overcoming the controversy.

Estonian LR: Money and Availability

Although LR are often created in academia, financed by the government, it is by far not the only way for finding money for LR.

The easiest way to make information (including LR) available to the public is via Internet. Our experience shows that committing oneself to making LR available via Internet is also the easiest way to find public and sponsor money. So one reason why Internet is suitable as a platform for LT is that there is more money available for that platform than for others.

A good example about mutual interests of a sponsor and LR creators can be found from Estonia. A world-famous sponsor George Soros considers free availability of information is to be of uttermost importance for an Open Society (Popper 1945). He helps to achieve this goal via Open Estonia Foundation[1], sponsored by himself. The foundation has given several grants to make LR (which can be viewed as a form of information) available in Internet.

These grants have actually also helped to create the LR, as Internet requires the material to be in a standardized form — before one could make the LR available to the public, a lot of correcting and cleaning had to be done.

In 1997 an Estonian national program for language technology was started. The aim of this program is to create LR for taxpayers' money and make them available to the public, via Internet again.

By now the list of available Estonian LR in Internet is a long one. There are several monolingual and bilingual dictionaries, text corpora and handbooks. One can also find a number of applications, usable via Internet and taking advantage of existing LR. The best starting points for looking for these applications and Estonian LR are:

- The home page of an Open Estonia Foundation spon-

[1]

- A page from the Estonia-Wide Web homepage:

  s/welcome.html
- The home page of the Department of Computational Linguistics: http://www.cl.ut.ee/
- The home page of the Institute of Estonian Language:

Internet Supporting LT

Kim and Choi (1996) describe Internet as the LT platform for Korean. The same idea is the basis for Estonian LT platform as well.

One can use Internet for downloading resources from a remote machine, then install them in the local machine and subsequently, use. This way Internet is just a medium for information exchange. We are here more interested in another way of taking advantage of Internet: to directly use LR which are designed for using on-line, without the possibility to download them. Such LR have the following characteristic features:

1. Standardized format. It is inevitable that the resources must have a uniform standard format: HTML, CGI etc. Unfortunately it is also true that some formats make it impossible to use the LR in any other way than was planned by the creators of the interface to the LR. E.g. some query formats for dictionaries do not allow for automatic queries, thus effectively excluding the dictionary from acting as a possible component of a translation aid.
2. On-going development. The resources are like a black box for the user: the interface stays the same, although what exactly goes on inside the box may change.
3. Availability. The resources stay under the control of their developers; this in turn makes the developers more willing to share them with the public.
4. Low cost. Using such LR is free or costs very little, as the aim of putting them online has usually been to promote some other product of the authors: a paper dictionary, a translation service etc.
5. No problems with copyright as the resources are only used (perhaps a little differently from the way the authors anticipated), but not copied.
6. New sources for LR. Material that the creators have put on-line appears to be LR without the creators being aware of it, e.g. on-line dictionaries, newspapers, books.

Integrating LR in Internet

Combining Dictionaries

Although dictionaries and programs may be available for using in Internet it doesn't mean one can really use them comfortably. E.g. a typical on-line dictionary assumes that the user questions it word by word, while the user is interested in using the dictionary as a true translator's aid.

If a user wants to look up the same word from several dictionaries, (s)he must submit several queries. If the queries have a standard format, making the queries might be a job of a program. The input of the program would be a word and the URLs of the various dictionaries; the output would be a compendium of the replies from all the queries.

Applications taking advantage of the uniformity of the query formats of various dictionaries are already appear-

than 200 on-line dictionaries (for English) and

-

tonian).

Comparing these two applications we note that as an an-

set of pointers to relevant dictionaries so that in order to get the information, the user must make more clicks on the

all the dictionaries at once.

Note that integrating the dictionaries does not mean one has to combine and harmonize the source texts. Integrating takes place via the query formats, the dictionaries themselves remaining "black boxes" for the integrating program.

This way one can minimize the effort in combining the dictionaries. Adding new ones to the combined set does not require re-structuring of the old ones. As integrating the dictionaries goes without any reference to the inner structure and source text, the authors of the dictionaries have no fear their intellectual rights are violated.

Combining Existing Modules

Figure 1 shows the result of spell-checking the home page of "LREC Workshop on Language Resources for European Minority Languages"[2].

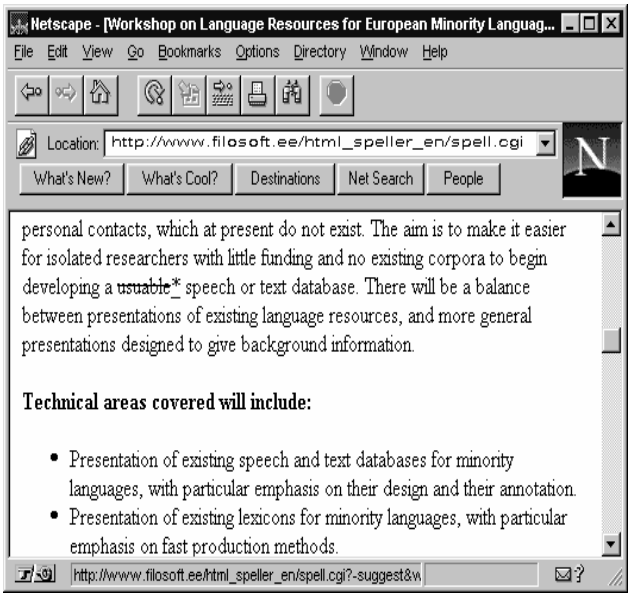

Figure 1: A speller for HTML-documents.

A spelling error has been found; it is indicated by a strike-through, followed by a hyperlink with an asterisk for the user to get suggestions for correct spelling. This gives an example of the simplest way of combining existing technologies: an existing spelling engine (in the given example an English one) and a program (based on Harvest[3]) for automatic downloading documents from Internet, based on their URLs. A feature like that is considered to be extremely handy for the end-user (Pedke 1998). The result is a speller that can be used for spelling

2

3

HTML-documents by their URLs, as well as ordinary text (which has to be input directly, e.g. by using cut-and-paste).
Instead of a spelling engine we may have the application built around some other LT module, e.g. a machine translation module, like in AltaVista Translation Service[4].

Linking LR to Documents

In the case of the speller and translation module, described above, the creator of the application had free hands in using and modifying the module in order to integrate it in his (her) application.
This is not always the case, however. A potentially widely usable LR are on-line dictionaries: they are usable via Internet, but not downloadable. Below we see how this kind of resource can be put to a new use.
Here is an example of a simple tool that takes advantage of a LR in Internet - an on-line dictionary.

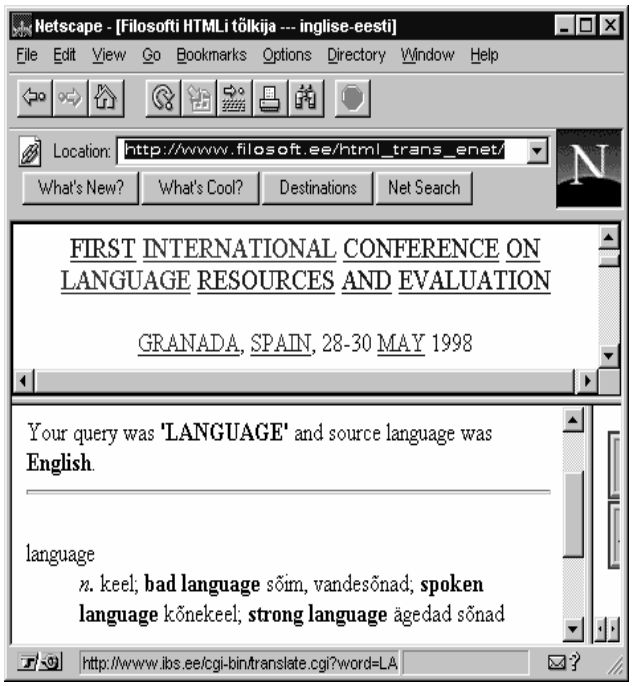

Figure 2: A simple translation-aid.

Figure 2 gives as an example an interface of a simple translation-aid. In the upper frame we see the home page of LREC. In the bottom frame we have the Estonian translation of the word "language", obtained by the user after clicking on LANGUAGE on the LREC home page. The translation comes from an English-Estonian dictionary in Internet[5]. The LREC home page in the upper frame is quite similar to the original[6], although not identical: every word is a link, and original links are converted to symbols like [->].
This is an example of how a dictionary, not intended to act as a LR originally, appears to be one, because it is available in the Internet and adheres to certain useful standards.

What's the mystery behind a tool like that, effectively integrating a LR into the end-user environment?
First, the application contains a program for automatic downloading documents from Internet, based on their URLs, like the speller and AltaVista Translation Service, described earlier. However, instead of an in-built module responsible for heavy language-related tasks, the application contains a module that converts a word into a hypertext link to the query form of an on-line dictionary. Figure 3 shows a portion of the source text of the converted homepage of LREC. Note the automatically generated links starting with "<a href=".

```
<p>
<font size=4 color="#800080">
                                    -
bin/translate.cgi?word=FIRST&language=Engli
sh" target="TransOutput">FIRST</a>
                 www.ibs.ee/cgi-
bin/translate.cgi?word=INTERNATIONAL&langua
ge=English"
target="TransOutput">INTERNATIONAL</a>
                                    -
bin/translate.cgi?word=CONFERENCE&language=
English"
target="TransOutput">CONFERENCE</a>
                              /cgi-
bin/translate.cgi?word=ON&language=English"
target="TransOutput">ON</a>
                                    -
bin/translate.cgi?word=LANGUAGE&language=En
glish" target="TransOutput">LANGUAGE</a>
```

Figure 3. Source text, generated by a translation aid.

This particular translation aid has the dictionary query form address and parameters hard-coded in it. However, it's obvious that it doesn't matter what the new hypertext link refers to, as long as it adheres to certain formal requirements. For example, the link may be to a query form of some other dictionary or even some text corpus. In the latter case, a click on the word would give examples of usage of the same word in the corpus.
So in principle it's possible that the user gives the dictionary name as a parameter, and (s)he gets a tool that does all the look-ups for him (her). Note that the requirements for the standardization that such a tool requires are widely met in the Internet: the query forms for various dictionaries are similar, and the texts are encoded in a uniform way.
Below we give a brief comparison of two systems that both allow a user to select his (hers) dictionary and use it as a reference point for the links, generated on the fly for the user's text: Hyperlinker by Filosoft[7] and Wordbot[8].
The output of both of the programs is a new hypertext where every word is a link to the LR the user selected; and a click on the word gives the user in a separate frame whatever information is available from the LR. The old links in the original are retained in a different form. The overall look of the on-the-fly generated hypertext is pretty close to the look of the original document.

4 -bin/translate?
5
6 .ugr.es/~rubio/elra.html
7
8 -js.html

Both programs accept as input an URL for the document to be processed. As an additional possibility, Hyperlinker accepts ordinary text as input.
Wordbot allows the user to select from a set of pre-defined dictionaries, not requiring the user to know the exact URL and parameters of the query form of the dictionary. Hyperlinker in contrast wants the user to provide the exact URL and parameters of the query (e.g.

Webster's dictionary).
To sum it up: Wordbot is more convenient to use in its limits while Hyperlinker provides more possibilities for the user.

Linking a Dictionary to a Corpus

A similar method like the one described above has been used for linking all the words in a corpus of Old Estonian with a dictionary of Old Estonian[9]. The corpus has not been linked with the dictionary "once and for all", but every time a user requests a text, the links are built anew. This way, a corpus and a dictionary, prepared separately initially, have been linked to each other in Internet.

Future Applications

As more and more LR find their way into the Internet, the number of applications making use of them inevitably rises, most notably the number of applications for new languages. However, besides similar applications, we may expect that completely new tools appear.
The ways for integrating LR via Internet, described above, represent only one simple method of taking advantage of existing technology. They do not cover a series of problems, solving of which would yield still more user-friendly applications.

1. Dictionaries usually assume that the input for their query forms must be a base form; but in the texts words are usually not in their base forms. Thus the module, responsible for transforming the words into hyperlinks, should include a lemmatizer.
2. It's often desirable to have multi-word expressions (like phrases) as hyperlinks, not only words.
3. It's rather common that we have a dictionary from language A to language B and a dictionary from language B to language C, but no dictionary from A to C. If we need a (however clumsy) translation from A to C, it would be natural to use a program that forwards the answers from one query to the next one, so that querying a word in language A would yield the corresponding words in language C.

Conclusions

Internet has a manifold influence on creating and using LR:

1. LR that are made available via Internet tend to be in a standard format.
2. It is relatively easy to get sponsor money for creating LR if the resources are intended to be available in the Internet.
3. Third parties often put material on-line that appears to be a LR without the creators being aware of it.
4. Completely new applications, taking advantage of on-line LR, become available (like translation aids).
5. The freedom to use LR in Internet is less hindered by the copyright problems.

Internet is not just a medium for transporting LR from one local point to another, but an environment allowing smooth integration of LR, e.g.

1. Various on-line dictionaries via their query forms.
2. Text documents with on-line LR via hyperlinks.
3. Text corpora with dictionaries via hyperlinks.

Although there exist a number of applications taking advantage of LR in Internet, there is still plenty of room for more sophisticated tools, as well as for similar applications for new languages.

References

Kim, S. & Choi, K-S. (1996). Korean Language Engineering: Current Status of the Information Platform. In *Proceedings of COLING 96, Vol. 2* (pp. 1049--1052). Copenhagen: Center for Sprogteknologi.

Pedke, T. (1998). Translation at the National Air Intelligence Center. *Language Today,* 6, 6--15.

Popper, K.R. (1945). *The Open Society and Its Enemies*. Routledge & Kegan Paul, London.

9