EASTIN-CL: A multilingual front-end to a database of Assistive Technology products

Gregor Thurmair
Linguatec
Munich
g.thurmair
@linguatec.de

Andrea Agnoletto
Fondazione Don
Gnocchi Onlus
Milano
aagnoletto
@dongnocchi.it

Valerio Gower
Fondazione Don
Gnocchi Onlus
Milano
vgower
@dongnocchi.it

Roberts Rozis
Tilde
Riga
roberts.rozis
@tilde.lv

Abstract

The document describes an application of language technology to improve the access to a database of Assistive Technology in the EASTIN-CL project. It focuses on engineering aspects of language technology integration. The paper describes the collection of a multilingual terminology database of the domain, and its use in multilingual and multimodal frontend components, especially the design, implementation and test of the query component. The system will be online for public web access under www.eastin.eu [1].

1 Context and Task

Access to information on Assistive Technologies (AT) is a key issue in social participation and eInclusion. The UN Convention on the Rights of Persons with Disabilities declares this as a fundamental right; all UN member states are obliged to comply with this Convention.

To support people with disabilities, the single states have organised web Portals which provide information about Assistive Technology products. Portals are visited by doctors, physiotherapists and other people in the domain.

The information in the AT domain is structured along the lines of the ISO 9999 standard ('*Assistive Products for Persons with Disability – Classification and terminology*'). This is a classification along functional aspects; AT databases group relevant products under each heading of this classification.

In 2005 the major European AT information providers joined in the European Assistive Technology Information Network (EASTIN). EASTIN provides a portal (www.eastin.eu) where people can access *all* databases of its national members simultaneously; the central server collects information on all products existing in one of the databases of the associated partners, so information seekers search on European level.

Fig. 1: Searching in national portals using ISO codes

However, variety of languages is still a barrier for easy access to AT information, although the portals provide at least an English translation in addition to the national languages.

To open the scope of this portal for additional user groups (end users), and to support people not familiar with the ISO 9999 classification, and speaking only their native language, a language technology front-end to the EASTIN portal was built in a project called EASTIN-CL. This front-end is supposed to be

- multilingual, i.e. users should forward information requests, and receive results, in their native language, and
- multimodal, i.e. users should be able to use a spoken channel of interaction in addition to the written one, cf. Fig. 2.

Supported languages are Danish, English, Estonian, German, Italian, Latvian, and Lithuanian.

A specific feature of the EASTIN portal is that search is not based on free text but on ISO 9999 codes.

[1] This project is partly funded by the European Commission, ICT-PSP no 250432. Partners are Linguatec, Tilde, Fondazione Don Gnocchi / SIVA, Institut der deutschen Wirtschaft / Rehadat, and Danish Centre for Assistive Technology / HMI.

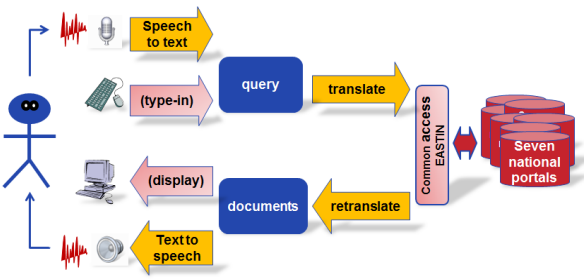

Fig. 2: EASTIN-CL frontend, EASTIN backend

So the approach is not *cross-lingual* as search terms need not to be translated, but *multilingual* as search terms in each single language point to ISO codes which can be used in search.

2 Master Term List

The first task was therefore to collect the concepts which form the AT domain, and to decide which ISO 9999 code they are linked to. The approach taken was to start with one 'master' language (English), and translate the resulting term list into other languages. Of course, most of the concepts are multiword terms.

2.1 Selection of Master Terms

The usual way to collect a selection of domain terms is to do corpus analysis, by collecting texts of the domain, and running term extraction tools. This way was tried first, and a list of about 120.000 term candidates was produced. However, this list was not usable, for several reasons:

- It was too large to be translated into seven languages with the resources of the project;
- Most of the high and medium frequency candidates were not specific enough to be included into the domain term list (i.e. assign a ISO 9999 code to them)
- In turn, many of the domain specific terms did not occur in the candidate list altogether.

So, the terms retrieved were not really suitable, and many good terms were not retrieved.

As a result, the approach was changed, and the domain terms were collected from existing descriptor lists: Many AT information providers, like Abledata, Rehadat etc.[2], offer key term lists for searchers, and so does the ISO 9999 classification itself. It was decided to base the domain terminology on such key terms, and to merge them into a common resource, resulting in a candidate list of about 17.000 terms.

Merging revealed that the different information providers had different strategies of key term denomination: Some presented them in plural form ('*wheelchairs*'), others in singular; some used US, others UK spelling, etc. As a result, the master list contained many pseudo-doublets. A cleanup step was needed based on principles like: use singular form as in paper dictionaries; use UK spelling ('*tyres for wheelchairs*' instead of '*tires for wheelchairs*'), use one term to describe one concept; i.e. split '*backrest (bath/shower)*' into two entries; use hyphens only for particles ('*dial-up*') or objects of participles ('*author-based*')

Even after cleanup, there is significant variance in the denominations. The final list contains about 12.700 concepts, with part-of-speech annotations.

2.2 Creation of the Domain Classification

All concepts should be linked to a domain ontology. In the case of AT, the domain is structured by the ISO 9999 classification[3], a three-level classification with about 800 nodes overall.

All terms of the master list were assigned one or several ISO codes; so the list forms a 'lightweight ontology' of the AT domain. It is expected that the term list will be fine-tuned during the test and use of the system.

2.3 Multilinguality

The task of creating the domain terminology was completed by translating the master term list into the seven languages of the EASTIN-CL partners.

Translations were carried out by domain experts, and in unclear cases the product databases could be consulted to find the best translation.

The resulting term list contains all 12.700 concepts, expressed in seven languages, about 90.000 terms altogether. This list was converted into the TBX standard, and is also offered for online access on the EASTIN-CL website.

3 Indexing and Search Preparation

3.1 Approach

In searches containing multiword terms, two indexing strategies are possible[4]: *pre-coordination* collects multiwords before searching; and *post-coordination* collects them afterwards (usually by *AND*-ing the single elements).

Nearly all search engines use post-coordination; however it can easily be seen that in multi- and crosslingual contexts, multiword terms must be

[2] www.abladata.com, www.rehadat.de, www.hmi.dk

[3] ISO 9999: 2011

[4] cf. Buder et al. 1990

recognised beforehand, as they may need a specific translation: if the parts of '*stuffed bag seat'* are each translated in isolation, the correct German translation into '*Sitzsack'* will not be found, and search results will suffer from this mistake[5]. In EASTIN-CL, the index contains multiwords, so pre-coordination is selected for indexing.

3.2 Index creation

Given the large variety in the term representations, the index terms must be considered as the target of a normalisation step, covering as many search term variants as possible.

The index in EASTIN-CL contains four fields: 1. the term in its *display form*, as it is presented to end users; 2. the term in its *normalised form;* 3. the single parts of the term as a sequence of *base forms* (lemmata), and 4. the *ISO* code(s) assigned to the term to find the real documents.

The representation of a multiword term as a list of lemmata requires lemmatisers and decomposers for the seven languages involved; tools by Linguatec and Tilde were used for this.

These tools had to be adapted to the AT domain (to decompose terms like '*Thorako|lumbal|orthese*', which would match a query for '*lumbale Orthese*').

3.3 Search preparation

The query processing component must analyse the query text; it needs language resources for this. In EASTIN-CL, two considerations influenced the design of these resources: 1. It is a *runtime* component, i.e. it is time and resource critical. 2. The EASTIN *target vocabulary* is limited, and basically a fixed set: Not *all* input words but only the words of the term list need to be recognized.

Therefore, a 'static lemmatiser' and a 'static' decomposer resource were implemented, whereby in a fixed lexical resource inflected forms point to their lemma, or word parts.

4 Search

Search in EASTIN-CL consists of three steps: query analysis and translation, search proper, and result retranslation.

4.1 Query analysis

Query Analysis must map a query input to an index term. The index term is annotated with ISO 9999 codes, pointing to groups of AT products.

[5] cf. [self-cite]

While the EASTIN portal is responsible for a distributed access to the national AT product databases, the frontend is responsible to produce ISO codes for searching, cf. Fig. 3.

For the language of the search, query analysis does tokenization, normalization, and lemmatisation and decomposition, by looking up the word form in the static resources. Finally, all index terms containing the single words of the query are retrieved as search candidates.

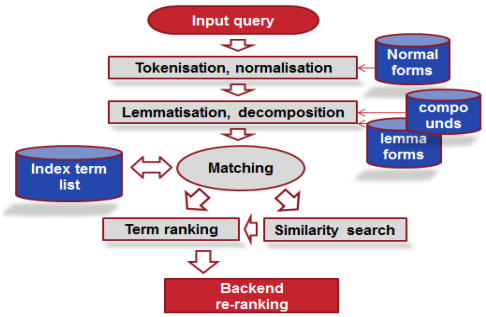

Fig. 3: Query Analysis. Resources needed: lemmatisers, decomposers, normalisers, in 7 languages

If no hit is found (typing errors), a fallback distance-based similarity search is used.

The final step of search is ranking the candidate terms. Ranking is based on the number of words in the query, the number of words of the index terms, and the number of matching terms. The terms with the highest overlap of matching terms are considered to be the best. The result is mapped on a 5-point scale, and the best ranked terms are returned with their ISO codes.

4.2 Searching

The search backend takes the candidate list of the query processing, and re-orders them as follows:

While the query processing takes care of the best matching *index term*, the main search intention is to find the best group of products, i.e. the best matching *ISO codes*.

Therefore the term list produced by the query is re-ranked based on term ranks and the ISO codes found, and the highest ranked ISO code (not necessarily the highest ranked term) is used for searching. This makes the system more robust. The search interface displays which term contributed to which ISO code (cf. Fig. 4).

To avoid a situation where users find no hits, the EASTIN portal offers additional search options, like search by navigation in the ISO classification; search for products ('*Tigges-Lumbal-orthese*') or manufacturers ('*All Terrain Wheelchairs Ltd*') with the search term in their name.

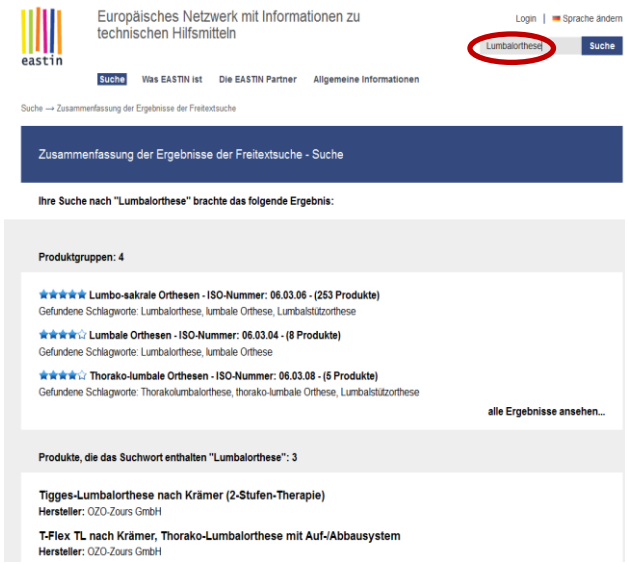

Fig. 4: Search in the portal for '*Lumbalorthese*': Search term produces a list of ISO codes, with the terms which retrieved them underneath. Ranking is given with stars.

4.3 Retranslation

Result of a search is a list of products, grouped under a given ISO code. The product descriptions in the national EASTIN databases are stored in the national language, and in English. The multilingual front-end now must re-translate the product descriptions into the query language. This translation is done on-the-fly: The EASTIN server accesses MT web services to translate the product descriptions. Both rule-based (Linguatec's '*Personal Translator'* English-> German/Italian) and SMT systems (Tilde's '*Let'sMT!*' platform, English->Baltic languages) are used. The MT systems were tuned for the AT domain, using the master term list and additional corpus data. Subject of translation are the textual parts of the product descriptions.

5 Evaluation

The objective of the evaluation was to find out:

If (end) users search for a certain AT product, which query terms do they really use? How good does the terminology provided match the search interests and search profiles of the users?

5.1 Evaluation approach

Two types of tests were designed:

The first test is a test on terminology. About 100 pictures of AT products were selected randomly, and put online, asking users to enter the terms they would use to search for the type of products depicted on them. Users can input queries, which are analysed to find out if the terms used point to the right product group. This procedure avoids to influence users by proposing terms, and allows to verify if the terminology provided by the EASTIN components is intuitive and of good coverage.

The second one is a test on usability. Users are given little tasks, and their interaction behaviour is evaluated with questionnaires: Does their search succeed? Which search tool do they use? Is MT of any help? etc.

5.2 Test Results

Tests of the term selection for pictures showed that users use terms which are recognised, and therefore lead to the right product group, in the majority of the cases (> 60%, with slight differences in the different languages);.this emphasizes the good coverage of the term list. Error analysis showed that this result can be further improved by adding synonyms to the term list.

Preliminary results of the usability tests, performed with about 60 external users, show a significant increase in the acceptance of the system, mainly due to the query functionality, but also to the machine translation and speech interaction components.

Overall, the language technology front-end components are considered to be a significant improvement in the accessibility of the Assistive Technology provided by the EASTIN portal.

References

Andrich R., 2011: Towards a global information network: the European Assistive Technology Information Network and the World Alliance of AT Information Providers, In: G.J. Gelderblom et al. (eds): Everyday technology for independence and care. pp. 190-197. IosPress

Buder, M., Rehfeld, W., Seeger, Th., eds., 1990: Grundlagen der praktischen Information und Dokumentation. Saur.

International standard ISO 9999:2011, Assistive Products for Persons with Disability – Classification and terminology

Lyhne Th., 2011: The Danish National Database on Assistive Technology, In: G.J. Gelderblom et al. (eds): Everyday technology for independence and care. pp. 205-213. IosPress

United Nations, 2007: The UN Convention on the Rights of People with Disabilities. In www.un.org/disabilities/

Winkelmann P., 2011. REHADAT: The German information system on assistive devices, In: G.J. Gelderblom et al. (eds): Everyday technology for independence and care. pp. 205-213. IosPress