When Automation Requires a Semantic Interface: Some Examples

Gerd Sjögren, Mats Granström, Henrik Andersen
Interverbum Technology, Copenhagen Translation
Stockholm/Sweden, Copenhagen/Denmark
{gerd.sjogren, mats.granstroem}@interverbumtech.com, henrik.andersen@copenhagentranslation.dk

Abstract

There are ambitious initiatives in sectors as diverse as Health, eCommerce and Process Industry to automate functions where semantic comprehension is a prerequisite for implementation. Without this comprehension, attempts to automate run the risk of becoming excessively time-consuming. The current presentation will focus on such cases. The main example is the Nordic Innovation-supported development project *Patient-Professional Communication Plugin (PPCP),* where *free-text* extracts from an Electronic Health Record (EHR) can be displayed, and difficult terms «translated» into everyday language, in real time. The interface allows the user to choose these «translations», originating from a termbase, in one or several of the five major Nordic languages as well as English, Arabic, Somali and Tigrinya. Another example, using the same termbase, shows how someone writing a medical text can be prompted to use either a standardised medical term, or its everyday equivalent, rather than «professional jargon».

The examples have implications for how easy it is for a novice to tackle «controlled language», typically in the form of a drop-down menu, in situations where the individual does not recognise or fully understand the available alternatives. The presented examples make use of a terminology tool working in the background.

Keywords: Electronic Health Record, automation, semantic interface, standard medical language, terminology, language tool, e-commerce, controlled language

1. Background – health management objectives

There is a strong push in all Nordic countries, the EU and elsewhere for patients to become engaged in the management of their own health by giving them access to their own EHRs through sites such as Kanta/FI, Journalen/SE and Journal fra sygehus/DA. The health benefits of this engagement are well recognised, but so are certain problems. Health records are traditionally written by professionals for themselves and other professionals, and the language used is – or has typically been - a mixture of professional terms, jargon, intra-clinic abbreviations and similar. Understanding what the EHR says is not even limited to the patient and his/her immediate caretaker; it is often equally difficult for any other non-specialist health care provider.

Examples of communication challenges:

- Will patients understand their medical record online?
- Do health information sites understand the patient?
- Do Watson and other decision support systems speak any language beyond English?
- Is there a way to help health care professional to use standard medical language without effort?

To solve these challenges, there is a need to link professional expressions with everyday language. Bridging the gap between professional and non-professional terminology is in fact essential in ensuring that e-health applications are user-oriented and contribute to a high quality of care. Helping patients better understand their medical records involves expanding acronyms and abbreviations, transforming idiolectic structures to standardised terms, and providing colloquial explanations without loss or distortion of information.

2. The project: developing and evaluating a pilot solution

The *Patient-Professional Communication Plugin* (PPCP)[1] project has created a technical infrastructure linking multilingual professional and non-professional language use. The platform is built through the integration of existing commercial tools for multilingual terminology management and text management. To provide a proof-of-concept for the technical solution, a pilot database specifically related to 'heart attack' has been developed. Heart attack is a potentially life-threatening diagnosis which is wide-spread and requires rapid and appropriate medical attention. Research shows that large patient categories may be underserved due to various reasons where communication is likely to play an important role.

The PPCP technical platform is intended for written and spoken digital exchanges. The fundamental approach of PPCP is to ***ensure accuracy and validity*** in the translation between professional and non-professional language. Until recently, machine translation systems such as Google Translate have primarily relied on statistical

[1] A Nordic Innovation-supported project formally named *Bridging the language gap : Creating a platform for patient-professional communication,* see https://www.nordicinnovation.org/programs/patient-professional-communication-plug-ppcp

frequency, which – as we know - often distorts results in unpredictable ways. The PPCP platform differs from this situation since all translations are pre-validated by professionals. It means PPCP may serve as the backbone for a very large number of uses, including the improvement of machine translation accuracy and validity.

3. Project display

A logged-in user first chooses his/her EHR base language, in which the medical record is written – see the upper right corner of the screen in figure 1. The User interface adapts its format to a variety of devices such as computer screens, tablets and smartphones.

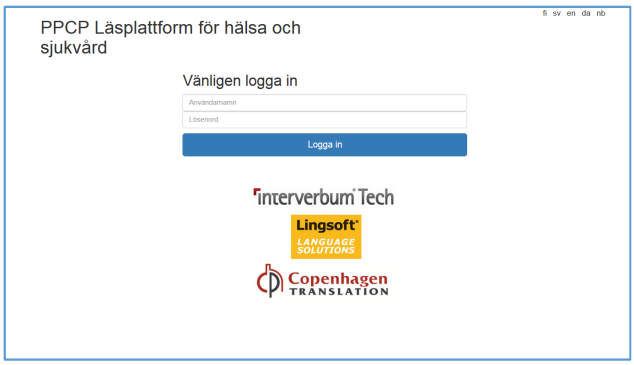

Figure 1: The PPCP login page

Once logged in, the user has a dropdown menu for choosing in which languages the explanations should be provided. The termbase behind the application contains professional and everyday language related to ´heart attack´ in Danish, Finnish, Norwegian Bokmål, Swedish and English. Some concepts and terms are also explained in Arabic, Icelandic, Northern Sami, Somali and Tigrinya. The difficult terms, where an explanation is available in the termbase, are highlighted in blue, and the user can see them one by one by pointing the curser to one of the highlighted words, see figure 2.

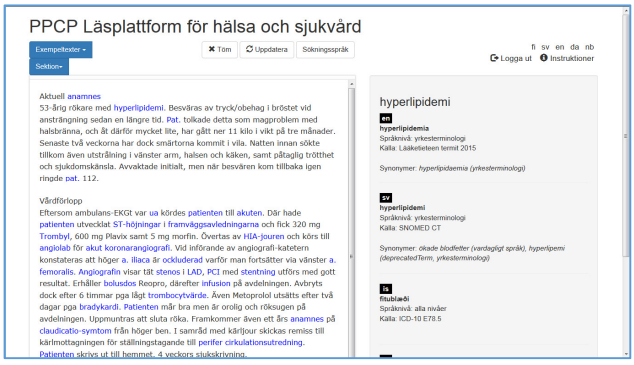

Figure 2: The EHR User interface

Since the content of the medical record is never altered by direct translation or interpretation, the intentions and tone of voice of its authors are never changed. We have learned this is an important point from a clinical point of view. All terminology and explanations in the database have been developed and validated by specialists. Anyone with a web browser and login information can work in the interface.

The solution is a proof-of-concept to show the principle of how a free-text health record can be made understandable. In addition, multimedia functions can be added to the EHR interface, so that images such as x-rays and magnetic resonance tests related to the patient's status, videos and sound files can be used to further explain concepts and terms. The underlying termbase tool allows this kind of multimedia support.

An online demonstration of PPCP is available on YouTube, https://youtu.be/PMAH4wS3Xpk. The demonstration is based on TermWeb, version 3. An updated version using TermWeb 4 will be soon be available.

4. Building the termbase

Since the PPCP project deals with medical terminology, there was a need to identify a relevant clinical subset of terms related to heart attacks. The idea was originally to identify the relevant clinical terms in Snomed CT and then translate them into all relevant languages, with the help of officially mapped results between Snomed CT and ICD-10. This turned out to be more difficult than anticipated due to formal requirements of Snomed International (*SI,* formerly IHTSDO), owner of the Snomed CT database. SI does not permit the development of professional terminology into languages of non-SI member countries. Consequently, we had to limit ourselves to already existing terms in Snomed CT and their officially mapped equivalents in ICD-10, meaning terms in Danish and Swedish. Existing ICD-10 terminology in other languages were then added. To extend the future usability of our heart-attack-related database, we decided to also include English, since it serves as the base language of both Snomed CT and ICD-10.

As a consequence, patient terms in languages such as Sami, Arabic etc. had to be mapped directly to the professional terms in the main language of the relevant country. Nynorsk was dropped on the advice of Norwegian specialists. Mapping minority language patient terms to the relevant main language concept was a risk mitigation strategy devised already in the original PPCP project plan. After considerable discussions and research, the collection of relevant terms for the database proceeded using the following four methods:

1) Introduction of clinical terms subsets (as mentioned above)
2) Monolingual term extraction from medical records in the main Nordic languages and English

3) Addition of "problem terms" provided by Reference group members such as heart associations
4) Supplementation of English term equivalents to Nordic language terms generated using methods 2 and 3

In practice, the work became part of a continuous workflow of gathering, evaluating, translating and validating concepts and terminology of both professional and patient language. The TermWeb tool permits searching for missing entries in the database through specially created filters. An advantage of this iterative process was that it opened for language-to-language translation rather than just professional-to-everyday language translation and vice versa. We also had to choose how to categorise terms, and ultimately came up with compulsory meta-tagging of the 'Language Level' of each term or explanation as either

- All (language) levels
- Everyday language
- Professional language, or
- Jargon and non-standard professional language

While adding terms, great care had to be taken to

1) Check that the term did not already exist. This process could be automated by using TermWeb's Batch Search Tool before import, or the Duplicate Check for single-concept editing
2) Check that a concept corresponding to the term had not previously been created in the database
3) When creating everyday language term equivalents to professional language terms, it sometimes became necessary to enter an explanation rather than a term

Extensive in-house and external resources have gone into developing the heart attack-related database. This includes the use of professional and medically trained translators for Arabic, Somali and Tigrinya. The number of entries for these languages is limited, however, especially in Somali and Tigrinya. The guidelines developed by the PPCP project group for building this kind of medical database can be requested from the authors.

5. Technical solution

The underlying analytical process in PPCP, shown schematically in ***Figure*** 3, is largely the same as it would be for other potential use cases, enabling further extensions of the platform.

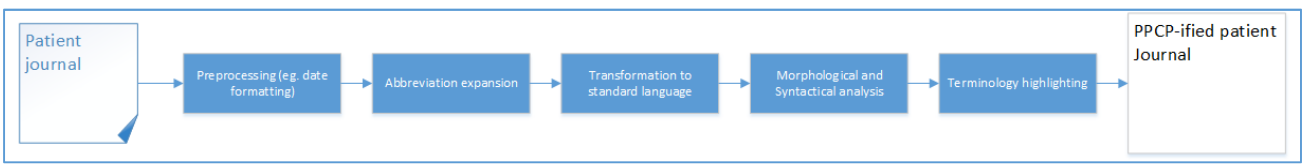

Figure 3: The analysis process for an EHR in the PPCP

A high-level overview of the underlying components in the platform is displayed in figure 4 below.

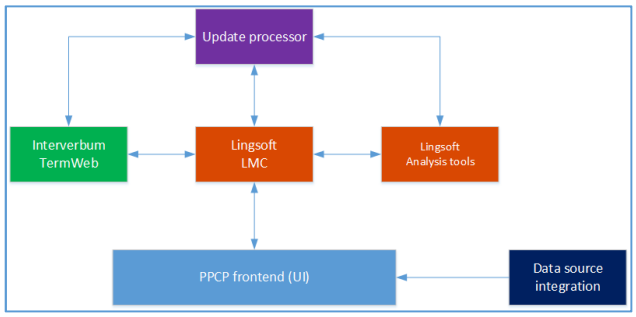

Figure 4: High-level component diagram

The technical infrastructure has been built by using the following existing software:

- Lingsoft's Language Management Central (LMC), a cloud service platform for language analysis. The platform allows authentication and user control for individual users or organisations, and an option for linguistic data collection to allow for analysis of user behaviour and semi-automated updates of the underlying language tools. In PPCP, there is total privacy for any entered information since no logs or copies are kept of the analysed text.
- Lingsoft language tools, the analysis components in LMC, including
  - Word and sentence analysis (entity name recognition)
  - Terminology and semantic analysis
  - Quality checking (spelling, grammar, stylistic rules)
  - Translation
- Interverbum's TermWeb, a terminology management tool, where the multilingual termbase is created and stored

6. Potential fields of application: health sector, eCommerce

- Proactively supporting clinicians in using standardised terminology. By integrating PPCP directly with **Electronic Health Records (EHR)** or writing tools, a physician could continue using his/her regular documentation routines (such as clinic-specific terms) while being prompted for the corresponding standardised terms. It implies dynamic interactive replacement of non-recommended terms. This would, in turn, facilitate automation of reporting requirements for national databases (such as heart or cancer registers) and improve safety aspects of the record. The need for costly terminology training and relearning for active health care providers would be reduced. Reduced costs of training would also be possible for professionals with a different native language.
- When electronic health records use standardised terminology, information in the records may more easily be integrated with clinical decision support systems. These range from regional or national recommendations regarding therapeutic action to the use of broad-based predictive systems such as IBM's Dr. Watson.
- Support for automation of communications between patients and health care providers by text or speech through contact points such as 1177.se, helsenorge.no and sundhed.dk. This communication could be urgent and very specific, where correct triage and documentation is needed.
- Powerful Dr. Watson-type systems that handle Big Data are typically monolingual. PPCP provides a potential avenue for broadening the scope of languages handled.
- Many excellent smartphone and tablet applications for the interactive management of medical conditions lack the 'language plugin' that would facilitate their use among the less educated, old, chronically ill or non-native language speakers. A language plugin would also open new geographic markets for these applications. PPCP can provide the necessary backbone for this multilingual use.

7. Using the termbase for controlled language purposes

In the PPCP interface, terms are recognised and explained in a "free text" context. This function has the potential for wide generic use, for example in e-commerce. PPCP has been demonstrated to individuals involved with Product Information Systems, where product descriptions for e-commerce platforms are automatically generated from a database. Different e-commerce platforms (Amazon, Ebay etc.) often require different terms for their respective product descriptions. This means "free text" analysis and subsequent conversion of certain terms to what is "correct" for the particular e-commerce platform could be of great interest.

8. Results and conclusions

The main result of the PPCP project is the availability of an accurate and reliable solution for helping patients better understand their own medical records. It is a solution applicable to all languages and language levels, using specialised and easy-to-understand wording. Results have been tested and evaluated by relevant patient groups and clinicians.

Additionally, the PPCP interface allows displaying how a health care professional writing free-text information can be prompted - in real-time - to use standardised rather than non-recommended terms, leveraging the same background functionality and database.

Among the deliverables in the project is also a method for gathering, translating and validating medically accurate terminology in several languages. This method, and the resulting termbase focused on heart attack-related terminology, form part of the project results.

In the PPCP project, only a limited number of languages – those most relevant in the Nordic countries – were included. With adequate semantic recognition support, there is no limit to how many or which languages could be supported by the database. The same generic functionality would also apply regardless of subject domain, given a good terminology database with similar (ISO TBX standard) data categories and values. This opens up a wide range of potential applications where controlled language is used, including e-commerce.

9. Acknowledgements

Nordic Innovation has provided financial support of the project. Special thanks to contributions from individual Reference Group members from several Nordic countries, and to the Danish Heart Association for their terminology support.