Getting to terms with terminology at Swedish public agencies

Magnus Merkel
Linköping University/Fodina
Linköping, Sweden
magnus.merkel@
{liu|fodina}.se

Henrik Nilsson
Terminologicentrum (TNC)
Stockholm, Sweden
henrik.nilsson@tnc.se

Abstract

This paper describes on-going work aimed at assisting public agencies in Sweden to conform to the new Swedish Language Act (passed in 2009). The Language Act highlights terminology as a key factor for a public agency, as well as a responsibility for a public agency to ensure that its terminology is made available, used and developed. Term-O-Stat is an action program to help public agencies to improve their terminological efforts. Term-O-Stat is divided into four distinct steps: 1) term inventory, 2) term classification, 3) conceptual analysis and term choice, and 4) term implementation. We describe the four steps and also experiences from the realization of step 1 and 2 at the Swedish Social Insurance Agency.

1 Introduction

A Language Act was passed in Sweden in July 2009. It contains a clause which clearly emphasizes that public agencies have a responsibility in making sure that Swedish terminology within their specific domain is "available, used and developed" (SFS 2009:600).

This means that there now is a clear legal incentive for public agencies to address terminological issues for their particular subject field. The specific terms that are currently being used within a public agency have often been developed over many years, and it is not uncommon that there is evidence of inconsistent term usage. For instance, a close scrutiny will usually reveal that one concept is denoted by a number of different terms on a website or in other public documents. Inconsistent term usage makes communication within a public government agency more difficult, and, furthermore, it can also make communication with citizens inefficient and confusing.

But, getting to terms with inconsistent and confusing term usage need not be that complicated. First of all, it is essential to investigate the actual term usage. Such an investigation can help to bring order in what terminology an organization is actually using. It is also important to try to specify the actual areas of responsibility for a public agency. The tax authorities have their specific responsibility to handle and maintain terminology for the area of taxation, and this differs from e.g. the Swedish Social Insurance Agency (Försäkringskassan). The latter will have to take the main responsibility for social insurance terminology. However, as terms from different areas are often used across public agencies, it is necessary to clarify who "owns" what terminology and also that terminologies will be shared across public agencies.

An effort to have public terminologies spread in Sweden was made by Terminologicentrum TNC (Swedish Centre for Terminology) in 2009 when they launched "Rikstermbanken" (www.rikstermbanken.se). Rikstermbanken is Sweden's national termbank on the web and holds over 77,000 term records spanning over a variety of domains. More than 150 organizations (most of them public agencies) have contributed to Rikstermbanken.

In many public agencies fragmented terminological resources are kept in Excel files or binders, but very few public agencies have systematically built a concept-oriented term database, and even fewer have integrated it in their writing environment.

2 Term-O-Stat

An action program called "Term-O-Stat" was launched in 2009 as an attempt to assist the public sector in Sweden to comply with the new Language Act, specifically directed to terminological issues. Term-O-Stat is constituted by the following four steps:

1. Term inventory
2. Term classification
3. Conceptual analysis and term choice
4. Term implementation

In short, step 1 concerns the collection and automatic analysis of documents in order to find what the actual term usage looks like. In step 2 and 3, the term candidates found in step 1 are processed further by their classification in subdomains, and the corresponding concepts are analysed and defined. In step 4, the results from steps 1–3 are implemented in a term database and a writing tool in order to integrate the established terminology into the normal writing and publishing workflow.

The overall program has been introduced to Swedish public agencies via seminars and on site visits.

In the following subsections the Term-O-Stat steps are explained in more detail.

2.1 Step 1: Term Inventory

In step 1, a document collection is analysed automatically, although some of the work involves manual inspection. The different phases of step 1 are the following:

1. Collection of suitable documents
2. Conversion from Word, Excel, Power Point, HTML and PDF to text
3. Grammatical analysis
4. Term extraction
5. Import to database
6. Filtering
7. Linguistic validation 1
8. Generation of synonym clusters
9. Linguistic validation 2
10. Cross-reference to internal linguistic resources (wordlists etc.)
11. Cross-reference to Rikstermbanken
12. Export to Excel sheet

The first phase involves a discussion with the agency in order to decide a suitable set of documents to use as input. This could result in all documents on the external website being selected, e.g. brochures, regulations, press releases, etc. The documentation volume can vary considerably between different agencies; for smaller agencies there may only be a couple of hundred thousand words, and for large agencies there may be several million words. The different file formats are then converted into plain text and sent to a grammatical analysis component. The grammatical and lexical analysis is made by Connexor's Machinese Syntax (Tapanainen and Järvinen (1997)). The analysis gives parts-of-speech information together with information on baseforms, morphological features as well as syntactic functions. After this step, term candidates are extracted; mainly noun phrases and verbs are extracted, but also syntactic function such as subject and object relations are utilized. The term candidates are then imported into a database and filtered (using stop word lists and syntactic patterns). All contexts for each term candidate are stored in the database and presented in an application (called TermViewer) where a linguist can validate the term candidates in context.

When the term candidates have been validated a search is made for synonyms among the candidates. This means that some term candidates are found to be possible synonyms to each other and therefore clustered together in a synonym set. The synonym clustering is made by string comparisons (for example the candidates "oral kirurgi" and "oralkirurgi" are clustered) and also with the use of Swedish synonym lexicons. The synonym clusters are generated automatically but inspected by a linguist for validation.

If the agency has internal word lists or lexicons, these are cross-referenced in order to find term candidates. A similar lookup is also made in Rikstermbanken and a reference is made for each term candidate that is also found there. At this point the data in the database is exported to an Excel sheet and presented to the agency. See fig 1.

C	D	E	G	H	I	J	
KonceptID	POS	Term	Frek	RTB	FKEN	FKBK	Exempel
#00042&	subst	oral kirurgi	5	oral kirurgi (Status:			Till--exempel kan åtgärd 491 bara debiteras av specialist
#00042&	subst	oralkirurgi	4				Om den som utför behandlingen är specialist eller utbilda
#00043&	adj	begärd	191				Du ändrar eller avbryter **begärd** föräldrapenning genom a **Begärd** ersättning Kontoförande bank är inte skyldig att pröva behörigheten Försäkringskassan har därför in en särskild skrivelse **beg** Försäkr Privat vårdgivare **Begärt**
#00043&	adj	beställd	6				En--del av en förklaring kan vara att fokus i slutrapporten Även beslut i massärenden utgör emellertid myndighetsu
#00044&	subst	F-skattesedel	2				Kopia på registreringsbevis från Skatteverket (**F-skattese**
#00044&	subst	Fskattsedel	4				Med innehav av F-skattsedel jämställs en skriftlig uppgift
#00044&	subst	F-skattsedel	116		corporate tax certificate		Det gäller även--om du är egen företagare med **F-skattse** **F-Skattsedel** **F-skattsedeln** utfärdas endast för ett år i taget . **F-skattsedelns** från och med datum är före eller samma

Fig. 1 Output data from step 1. Three synonym clusters with term candidates are shown with frequency data, cross-references and sample context.

2.2 Step 2: Term Classification

In step 2 of Term-O-Stat, the term candidates found in step 1 are used as input. Terminologists inspect all term candidates and classify them into different groups:

1. Terms specific to the public agency
2. Terms common in the public sector
3. ”General” terms
4. Non-terms
5. Names

Group 1 is the most important category for any given public agency. Terms classified as belonging to that category are terms that constitute “their own terminology” and thus the agency’s area of responsibility. The second group contains terms that are not unique for the agency but which are also relevant for other public agencies. Groups 3–4 will contain term candidates that are deemed not important enough to investigate further. These candidates can be terms of a more general character and words that superficially look like terms but belong to general language. It has proven useful to separate names in a special category (group 5). The classification of the term candidates is made by using the database produced in step 1 and by using the GUI interface TermViewer, which allows several users to access the database simultaneously. Terms from group 1 are further subclassified; the public agency in question may have its own classification system that can be applied here.

2.3 Step 3: Conceptual Analysis and Term Choice

In step 3, the terminologists continue to work with the terms in group 1 (terms specific to the public agency), together with experts from the public agency. The work here is more of traditional terminology work where concept clusters are analyzed and described in concept systems and then defined. The main difference from traditional terminology work procedures lies in that the starting point is a fuller inventory and categorized terms (in synonym clusters) that emanate from a large amount of public agency documents. The objective is to come to a consensus about the concepts, how they should be defined and what terms should be used to denote them.

An important aspect of step 3 is also to work with term status. If a concept is denoted by several terms, one term may be “recommended”, another “admitted” while three other terms could be classified as “deprecated”. This is very important and useful information when the terminology is to be “put to use”. The status indication of a term is a prerequisite for the integration of the terminology into the existing writing tools of the authoring and publishing environment (see step 4). Although agency-specific terms (group 1) are in focus in step 3, terms from group 2 could also become important, and this would entail the co-operation with other public agencies.

2.4 Step 4; Term Implementation

In step 4, the objective is to integrate the results from the earlier steps into the authoring and publishing environment. This is actually one way of complying with the Swedish Language Act that states that the terminology within the subject domain of a public agency also should be used, e.g. in the documents and website created by the agency.

It is not enough to publish a web page on the intranet listing all terms alphabetically to promote “usage”. It is of course better than nothing, but the optimal solution would be if the terminology could be integrated and embedded in the writing tools (word processors, presentation or web authoring tools, etc.) and used in the manner of ordinary spellcheckers and grammar checkers in applications like Microsoft Office.

Remembering terminology suggestions and detailed writing recommendations is extremely difficult for authors. Many public agencies in Sweden conduct regular training on writing and the use of proper terminology, but it is still hard to spread information on newly changed terminology and new policies. If it were possible to make changes to a central language server such changes could be made available directly through a language checking plug-in programme for the standard word processor.

One example where a central language server combined with language checking plug-in clients for various applications is acrolinx IQ. In acrolinx IQ it is possible to check documents for terminology, spelling as well as grammar and style rules. In other words, from a terminological point of view one can

- store and administer terms in an integrated term database
- highlight terms in documents that are admitted by the term database
- mark deprecated terms when such are used and propose a recommended term instead
- manage different term sets for different text types, users and domains within an organization

- extract new term candidates from existing documentation

The language checking can be performed by the author from the plug-in client or applied as a batch checking process on a set of documents.

3 Term-O-Stat so far …

Term-O-Stat has been active for approximately one year and during that time three open seminars have been held where more than 25 public agencies have attended. Step 1 and 2 have been implemented at Försäkringskassan (Swedish Social Insurance Agency). At Försäkringskassan, around 2,000 documents were processed in step 1, resulting in 17,000 term candidates that were fed into step 2. In step 2, the term candidates were distributed over the category groups in the following way:

- Terms specific to the public agency (2,628)
- Terms common in the public sector (2,320)
- ”General” terms (6,235)
- Non-terms (4,618)
- Names (726)

The first group, with terms specific to Försäkringskassan’s area of responsibility, was divided into eleven subareas, e.g. administration, housing, dental care, immigration, disease, etc.

4 Conclusions

The first Term-O-Stat project showed that the agency-specific terminology is much more complex than one could have expected, and also that there may be a considerably higher degree of inconsistency in how terms are used in practice. By using existing term lists in the inspection, it is possible to compare these to the actual usage in the analyzed document set. At Försäkringskassan, it was discovered that a number of terms specified in a rather small termbank were not used at all in any of the documents on the external website. This does not have to mean that they are unimportant, instead it may reflect the fact that the termbank focused on concepts that are not used in external communication.

So far, we have only dealt with monolingual term extraction. Bilingual and multilingual material exists but usually makes up only a fraction of the information published in Swedish. If parallel texts were available it would be possible to do bilingual term extraction and find terminological inconsistencies in both the source and target texts.

Automatic term extraction methods, filtering techniques, database technology and the performance of modern computers open up new exciting possibilities for making terminology projects much more efficient. On the other hand, terminology work requires access to domain experts, in this case experts at the public agency that has the in-depth knowledge of their subject area. The participation of the public agency representatives will be necessary for the following activities:

- Search and select documents that form the input data.
- Assist in thee categorization and clustering of term candidates (classification systems, clustering criteria).
- Participate in concept analysis, definition writing and term selection.
- Assist in the publishing of the material internally and externally,
- Train users in using new tools in the internal authoring environment.

The exact time that is required for these activities varies from agency to agency. A successful end result will to a large extent depend on how much time the agency can devote to the project, especially to step 3.

By combining automatic methods from language technology with manual validation and categorization, Term-O-Stat has shown that it is possible to get an overview of terminology usage that would have been practically impossible to acquire using only traditional terminological methods.

References

Foo, Jody & Merkel, Magnus. 2010. Computer aided term bank creation and standardization: Building standardized term banks through automated term extraction and advanced editing tools. In Marcel Thelen & Frieda Steurs (red.), Terminology in Everyday Life, (pp. 163–180). John Benjamins Publishing Company. Amsterdam.

Tapanainen, Pasi & Järvinen, Timo. 1997. A non-projective dependency parser. In Proceedings of the 5th Conference on Applied Natural Language-Processing, pp. 64–71, Washington, D.C., April. Association for Computational Linguistics.

Språklag (Language Act), Svensk författningssamling (SFS 2009:600). http://www.riksdagen.se/webbnav/index.aspx?nid=3911&bet=2009:600.