Detection of synonymy links between terms: 
experiment and results

Thierry Hamon and Adeline Nazarenko
Laboratoire d'Informatique de Paris-Nord
Université Paris-Nord, avenue J.B. Clément
93430 Villetaneuse, France
firstname.lastname@lipn.univ-paris13.fr

This paper presents a new approach for the evaluation of the detection of synonymy relations between terms. Our goal is to help the terminology structuring. This approach exploits the synonymy relationships which have been extracted from lexical resources to infer synonymy links between complex terms. The inferred links are then validated by an human expert in the context of a terminological application. In a previous evaluation on documents dealing with electric power plant, the expert has underlined that the most important point is to increase the recall even if the precision is low and if some links are mistyped. This paper reports new experiments which help to understand how this synonymy detection approach is to be used. Various lexical resources – from general language dictionary to very specialized semantic information – are exploited and compared as bootstrapping knowledge. Results show the complementary of the different sources.

The first evaluation relied on traditional recall and precision measures. However, those scores do not reflect the usefulness of the inferred links for the terminology structuring. From the terminologist's point of view, erroneous links are quick to eliminated. They may even suggest good ones. Above all, the system points out relations between terms which are generally not found manually. We thus aim at proposing a new evaluation criteria which better reflects the expert's and terminologist's point of view in the application context. This score points out the quality of the results and the validation cost rather than the proportion of validated links. We have designed an evaluation score which takes into account the productivity of the dictionary links. It can be viewed as a normalization of the precision.
1. Introduction

Document Consulting Systems require a terminology to help the user’s navigation through the document. Structured terminologies enhance the document access by providing information about the terms and their semantic links. For instance, it is useful to know that the terms matériau électrique (electric equipment) and équipement électrique (electrical fittings) or marche normale (normal running) and bon fonctionnement (right working) are synonymous if one wants to identify the various sections of the document that deal with electric equipment.

Several tools are available to help terminologists in the terminology structuring tasks. We focus on the detection of synonymy links between terms. We rely on a rule-based approach using general lexical resources, i.e. synonymy links between words extracted from a general language dictionary. The rules have been tested and tuned on different corpora of various size and domain.

We report here a new experiment which results from a collaboration with French electricity company (EDF). This industrial context and the interaction with users helped us to understand how the acquisition process should be bootstrapped and evaluated in the context of realistic terminological applications.

The results have been analyzed by the terminologist in charge of the terminology construction process. Various lexical resources have been exploited as bootstrapping knowledge from general language dictionary to very specialized semantic information. Results show the complementarity and the usefulness of the different sources. The evaluation phase showed that the traditional precision and recall are difficult to measure and do not take into account the terminologist’s usefulness judgment. We propose a new evaluation measure for terminology structuring tasks.

The following section presents the terminology building process in the context of a Consulting System and describes our approach for the detection of synonymy links. The acquisition of synonymy links using a general dictionary is first tested. The section reports this experiment analyses the results. This experiment then leads us to study the contribution of various lexical sources to the terminology structuring (section ). The section addresses the evaluation problem. A new evaluation measure is proposed, which better reflects the help of our system and of terminological tools to the terminology structuring.
Detection of synonymy links between terms: experiment and results

2. Enhancing a Document Consulting System with a structured terminology

2.1. The Document Consulting task

The increase of technical documentation leads companies to build document navigation systems. The Technical Documentation Consulting System (TDCS) developed at the French electricity company Electricité de France (EDF) by Gros and Assadi (1997) provides an hypertextual access to information contained into technical documents. It offers various access modes:

- a traditional table of content;
- a full-text keyword access as in information retrieval systems;
- an index of the activity modeling the user’s task;
- an index of the domain with synonymy and hyponymy links between entries and sub-entries (the part of the terminology of the domain which is relevant for the document).

The indexes are also used to expand or specify the user queries.

Consultation systems differ from hypertextual systems such as SNITCH (Mayfield and Nicholas 1992) which deals with a document base composed of short texts. SNITCH automatically and dynamically generates a semantic representation of a document in the form of a semantic network whose nodes are the document terms. Consultation systems do not process the whole document, they rather focus on the queries, helping the users to expand or specify their keyword list and proposing various access modes. This latter approach, which is not as costly as the former one and which is less sensitive to errors is well-suited for large technical documents such as a technical service manual.

 Whereas information retrieval systems aim at finding relevant documents from a large corpus of various theme for a given query, consultation systems usually deal with a single document of few hundreds of pages (less than one million words). In the former systems, the relevancy of documents is based on statistical scores or general semantic networks such as WordNet (Richardson et al. 1994) whereas consulting systems use terminologies i.e. domain dependent resources, to compute relevancy measures.
2.2. Terminology structuring

The integration of a new document in a consultation system is efficient if there exists a terminology of the corresponding domain. As such terminologies are seldom available for specific technical domains, they must be built automatically.

However, the terminology building process cannot be fully automated. Identifying of the terms of a technical domain is a matter of word usage and expert agreement. The detection of term links requires a thorough knowledge of the underlying concepts. There also exists various types of terminologies (Srinivasan 1992). The aim of the French electricity company EDF is therefore to develop a tool that helps the terminology building task. The system suggests a first selection of terms and term relations. The terminologists then correct and/or extends this sketch. Even if they have to go back to the document to check some specific points, they can avoid reading the full document.

The overall architecture of such terminology building tools is usually divided in two subsystems. The terminology extraction tools acquire terms from a corpus and build a term list. The structuring tools build a semantic network out of this list, by adding semantic links between the terms. Both the terms and the term relations require to be controlled by the expert. This two-step computer-aided process is adopted in Termight (Dagan and Church 1994) in the case of multilingual terminologies, for instance. The architecture of the Technical Document Consulting System is similar. A terminology extraction software (LEXTER (Bourigault 1992)) first automatically extracts candidate terms from documents. The terms are either simple or complex if they are composed of several words. The terms form a syntactic network which is then enriched with semantic relations during the structuring phase.

Traditionally, a terminology has different types of semantic relations (Srinivasan 1992: 162): hierarchical relations (is-a) which link a term to a more specific/generic one, equivalence relations (synonymy) and non-hierarchical relations (see-also), among which are the preference relations connecting a controlled term to its less accepted variants. The accuracy and the number of these relations vary from a terminology to another.

There is no general method for terminology structuring. Several complementary tools have been designed, each one dealing with a specific type of semantic relations. They respectively provide the terminologist with

- hypernymy, the hierarchical relation that traditionally structures the terminologies and allows for generalization inferences (Hearst 1992,
Detection of synonymy links between terms: experiment and results

Morin 1998): substances > bromine; substances > hydrogen or cation > calcium (were > stands for "is the hyperonym of");
- term variants, the phrases that are semantically equivalent to a term (Jacquemin 1996): knowing these variants helps the terminologist to choose the controlled term: gene expression vs. expression of gene;
- conceptual word classes or groupings (Habert et al. 1996, Assadi 1997, Pereira et al. 1993), sets of words which require to be hierarchically organized: {225 kV double circuit line, overhead line, 225 kV underground line};
- causal relations, which are especially important in technical domains where diagnosis, repair, planning, physical action are involved (Garcia 1998): to_create(electric charging increase, electric network expansion);
- theme and domain clues, which are also used to disambiguate terms: in the EDF Thesaurus, the entry 'behavior' is tagged [psychology];
- etc.

2.3. Acquisition of synonymy links between terms

Our synonymy definition is close to that of WordNet (Miller et al. 1993). It is a simplified one. We consider synonymy as a Boolean rather than as a scaling property as in Cruse (1986). We define a "contextual synonymy": two terms X and Y are synonymous relatively to a context C if both terms are syntactically identical and semantically substitutable in the context C.

We assume that the semantics and the synonymy of the complex candidate terms are compositional. This working hypothesis is discussed in Hamon and Nazarenko (1999). We assume that complex terms are composed of a head and an expansion. Two terms are considered as synonymous if their components are identical or synonymous.

We designed three rules to detect synonymy relations between candidate terms. Given the complex candidate terms $\text{CCT}_1 = (T_1, E_1)$ and $\text{CCT}_2 = (T_2, E_2)$, and $\text{syn}(\text{CT}_1, \text{CT}_2)$ a synonymy relation between the candidate terms $\text{CT}_1$ and $\text{CT}_2$, following inference rules are used:

$R_1: T_1 = T_2 \land \text{syn}(E_1, E_2) \supset \text{syn}(\text{CCT}_1, \text{CCT}_2)$

$R_2: E_1 = E_2 \land \text{syn}(T_1, T_2) \supset \text{syn}(\text{CCT}_1, \text{CCT}_2)$

$R_3: \text{syn}(T_1, T_2) \land \text{syn}(E_1, E_2) \supset \text{syn}(\text{CCT}_1, \text{CCT}_2)$
This means that, considering two candidate terms, if one of the following conditions is met, a synonymy link is added to the terminological network:

- the heads are identical and the expansions are synonymous (collecteur général (general collector) / collecteur commun (common collector)), \( R_1 \);
- the heads are synonymous and the expansions are identical (matériau électrique (electric equipment) / équipement électrique (electrical fittings)), \( R_2 \);
- the heads are synonymous and the expansions are synonymous (marche normale (normal running) / bon fonctionnement (right working)), \( R_3 \);

3. **Experiment**

3.1. **Experimental data**

3.1.1. **Working corpora**

In the context of our collaboration with EDF, we applied the proposed rules on the DSE corpus (DSE2 in the following). This 160 000 words corpus is composed of Elementary System Files (Dossiers de Système Elémentaire, henceforth DSE) and deals with the operations on nuclear power plants. This corpus has been analyzed by LEXTER (Bourigault 1992) which extracts 17 675 candidate terms: 4 171 simple candidate terms (2 865 nouns and 1 306 adjectives) and 13 504 complex candidate terms i.e. noun phrases. Each complex candidate term (ligne d'alimentation, supply line) is analyzed into a head (ligne, line) and an expansion (alimentation, supply). It is part of a syntactic network (cf. Figure 1).

Reported results are compared with a first DSE corpus (DSE1 in the following) on which we initially tuned the rules (Hamon et al. 1998). This 200 000 word corpus deals with the same domain. It has been analyzed by LEXTER: 12 043 candidate terms have been extracted: 3 428 simple candidate terms (2 831 nouns and 597 adjectives) and 8 615 complex candidate terms.

The synonymy detection process has also been applied on the Menelas corpus (Zweigenbaum 1994), a collection of documents dealing with coronary diseases.
Detection of synonymy links between terms: experiment and results

LEXTER analyzed this 85,000 words corpus and 13,233 candidate terms have been extracted (2,675 nouns, 1,006 adjectives and 9,552 noun phrases).

Results of the term extraction for two different corpora are summarized in the Table 1.

Table 1: Results of the candidate term extraction

<table>
<thead>
<tr>
<th>Corpus</th>
<th>DSE�</th>
<th>DSE₂</th>
<th>Menelas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Candidate terms</td>
<td>12,047</td>
<td>17,675</td>
<td>13,233</td>
</tr>
<tr>
<td>Simple candidate terms</td>
<td>3,428</td>
<td>4,171</td>
<td>3,681</td>
</tr>
<tr>
<td>Nouns</td>
<td>2,831</td>
<td>2,865</td>
<td>2,675</td>
</tr>
<tr>
<td>Adjectives</td>
<td>597</td>
<td>1,306</td>
<td>1,006</td>
</tr>
<tr>
<td>Complex candidate terms</td>
<td>8,615</td>
<td>13,504</td>
<td>9,552</td>
</tr>
</tbody>
</table>

For the reported experiment, we used a French general dictionary Le Robert supplied by the Institut National de la Langue Française (INaLF). The dictionary provides synonyms distributed among the different senses of each word entry (see Figure 2). It is exploited as a synonym dictionary. We choose a standard and widely available – although imperfect – dictionary so as to carry out our experiment in realistic conditions. The semantics of the extracted relations is often fuzzy and sometimes reflects mere see-also relations that one would hardly consider as synonymy. For instance, our dictionary extraction gives temps (time)/température (temperature) as a synonymous pair whereas the synonymy holds for a very specific meaning of both words (in the meteorological context). As we will see below, such weak dictionary relations generate erroneous links in the final terminology. This is partly due to the fact that we do not take into account the various sense distinctions of the dictionary entries, as in Ploux and Victorri (1998).
Although "contextual synonymy" is a transitive relation, we cannot consider the transitivity property, because the dictionary synonymy relations are context-free. The application of this property would produce a great deal of errors. However, most of the relevant links which can be inferred by transitivity already exist in the dictionary.

3.2. Validation protocole
The validation protocol differs from that of our former experiment (Hamon et al. 1998) in the presentation of the results and in the possibilities given to the expert.

A domain expert who is also a terminologist has validated the inferred links in the context of an access to the information contained in documents.

The interface allows to modify the relation type of each link. Although we aim at detecting synonymy relations between terms, other relations are inferred.

Moreover, whereas the rules are applied on lemmatized terms, we choose to present the term inflected forms of the terms for the sake of validation comfort.

Two presentations have been used for the validation of the inferred links. Besides the connected components which give an overall view of the links, we propose to structure the results in families.

Indeed, we have noted in the validation of the DSE1 corpus (Hamon et al. 1998) that all the links inferred from some dictionary word couple are rejected. This is the case for capacité (capacity) / puissance (power):

Figure 2: Example of a word entry from the dictionary Le Robert

modèle (model) <1> canon (canon), étalon (standard), exemplaire (copy), exemple (example), plan (plan)
<2> sujet (subject), maquette (maquette)
<3> héros (hero), type (type)
<4> échantillon (sample), spécimen (sample)
<5> standard (standard), type (type), prototype (prototype)
<6> maquette (model)
<7> gabarit (size), moule (mould), patron (pattern)
Detection of synonymy links between terms: experiment and results

We chose to group inferred links in families. A family is a set of links inferred from the same dictionary link. The links inferred with the third rule belong to two families corresponding to two initial links. For instance, the link *capacité faible* (low capacity) / *volume petit* (small volume) belongs to the families inferred from *capacité* (capacity) / *volume* (volume) and from *faible* (low) / *petit* (small).

This results presentation therefore helps the validation phase. The terminologists can easily reject a whole family if they consider that the dictionary link is erroneous for a given corpus.

3.3. Results

3.3.1. Overall results

We applied the rules on the DSE2 corpus. The inferred links have been validated by an expert of the domain according to the protocol presented above. 590 links have been inferred between complex candidate terms. The expert agreed on 33.72% of these links *i.e.* 199 links (cf. Table 2). Synonymy links represent 50.75% (*i.e.* 101 links) of the validated links *catégorie sismique* (seismic category) / *classe sismique* (seismic class).

As expected, see-also relations are numerous. They represent 42.22% of the links (*i.e.* 84 links) *échelle logarithmique* (logarithmic scale) / *gamme logarithmique* (logarithmic range) (cf. Table 3).

The expert identified hyponymy relations *phénomène naturel* (natural phenomenon) / *phénomène physique* (physical phenomenon) or meronymy relation *équipement REA* (REA fittings) / *matériel REA* (REA equipment).

The links validated as antonymy relations are often due to biases of lemmatization and presentation. The term lemmatization modifies some terms with negation mark. For instance, *matériels non statiques* (non-static equipment) is lemmatized as *matériel statique* (static equipment). Therefore, we infer the antonymy link *matériels non statiques* (non-static equipment) / *équipement statique* (static fittings).
Table 2: Results and validation of the links

<table>
<thead>
<tr>
<th></th>
<th>Number</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validated links</td>
<td>199</td>
<td>33.72%</td>
</tr>
<tr>
<td>Invalidated links</td>
<td>391</td>
<td>66.28%</td>
</tr>
<tr>
<td>Total</td>
<td>590</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 3: Type of the validated links

<table>
<thead>
<tr>
<th>Type of relation</th>
<th>Number of links</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synonymy</td>
<td>101</td>
<td>50.75%</td>
</tr>
<tr>
<td>See-Also</td>
<td>84</td>
<td>42.22%</td>
</tr>
<tr>
<td>Hyperonymy</td>
<td>11</td>
<td>5.53%</td>
</tr>
<tr>
<td>Meronymy</td>
<td>1</td>
<td>0.5%</td>
</tr>
<tr>
<td>Antonymy</td>
<td>2</td>
<td>1%</td>
</tr>
<tr>
<td>Total</td>
<td>199</td>
<td>100%</td>
</tr>
</tbody>
</table>

The distribution of the links varies from a rule to another (cf. Table 4). The inference rule 2 is the most productive whatever the relation, except for the hyperonymy. This shows the complementarity of our approach compared with the terminologist work. 67% of the synonymy links proposed by our approach have synonymous head while terminologists generally build links corresponding to the two other rules. The rule 3 is the less productive rule as we have already noted in Hamon et al. (1998). However, such links are generally not found by the terminologists.

Table 4: Distribution of the links according to the inference rules

<table>
<thead>
<tr>
<th>Type of relation</th>
<th>Synonymy</th>
<th>See-Also</th>
<th>Hyperonymy</th>
<th>Meronymy</th>
<th>Antonymy</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule 1</td>
<td>28</td>
<td>17</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>53</td>
</tr>
<tr>
<td>Rule 2</td>
<td>67</td>
<td>61</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>134</td>
</tr>
<tr>
<td>Rule 3</td>
<td>6</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>12</td>
</tr>
</tbody>
</table>

3.3.2. Analysis of the result structure
The distribution of the family size is identical for the three corpora (cf. Table 5). Most of the families are composed of one link (more than 2/3). The families with more than two links represent less than 1/10 of the families.
Detection of synonymy links between terms: experiment and result

Table 5: Distribution of the families according to their size

<table>
<thead>
<tr>
<th>Size</th>
<th>DSE₁ corpus</th>
<th>DSE₂ corpus</th>
<th>Menelas corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>179 (67.29%)</td>
<td>314 (67.38%)</td>
<td>286 (71.5%)</td>
</tr>
<tr>
<td>2</td>
<td>40 (15.03%)</td>
<td>84 (18.03%)</td>
<td>60 (15%)</td>
</tr>
<tr>
<td>3</td>
<td>19 (7.14%)</td>
<td>32 (6.87%)</td>
<td>20 (5%)</td>
</tr>
<tr>
<td>4</td>
<td>13 (4.88%)</td>
<td>16 (3.43%)</td>
<td>7 (1.75%)</td>
</tr>
<tr>
<td>5</td>
<td>1 (0.38%)</td>
<td>6 (1.3%)</td>
<td>11 (2.75%)</td>
</tr>
<tr>
<td>6</td>
<td>7 (2.64%)</td>
<td>7 (1.5%)</td>
<td>4 (1%)</td>
</tr>
<tr>
<td>7</td>
<td>3 (1.12%)</td>
<td>2 (0.43%)</td>
<td>5 (1.25%)</td>
</tr>
<tr>
<td>8-21</td>
<td>4 (1.52%)</td>
<td>5 (1.06%)</td>
<td>7 (1.75%)</td>
</tr>
<tr>
<td>Total</td>
<td>266 (100%)</td>
<td>466 (100%)</td>
<td>400 (100%)</td>
</tr>
</tbody>
</table>

We compute the homogeneity of each family $i$ which is defined as:

$$Homogeneity_i = \frac{|NbVal_i - NBErrors_i|}{NbLinks_i}$$

where $NbLinks_i$ is the size of the family, $NbVal_i$ is the number of validated links and $NBErrors_i$ is the number of erroneous links.

The homogeneity of a family is 1 if all the links are either validated or invalidated. The ratio is null when half of links are validated.

Analysis of the results of the DSE₂ corpus shows that the validation is homogeneous for 90% of the families (cf. Table 6): 69.52% of the families have been entirely rejected while 21.67% have been entirely accepted (cf. Table 7). We note a similar distribution for the DSE₁ corpus.
Table 6: Homogeneity of the families according to their validation

<table>
<thead>
<tr>
<th>Homogeneity</th>
<th>Number of families</th>
<th>Percentage</th>
<th>Number of families</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>9</td>
<td>3.38</td>
<td>17</td>
<td>3.65</td>
</tr>
<tr>
<td>0.14</td>
<td>1</td>
<td>0.38</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>0.2</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>0.43</td>
</tr>
<tr>
<td>0.333</td>
<td>6</td>
<td>2.25</td>
<td>16</td>
<td>3.44</td>
</tr>
<tr>
<td>0.42</td>
<td>1</td>
<td>0.38</td>
<td>1</td>
<td>0.21</td>
</tr>
<tr>
<td>0.5</td>
<td>4</td>
<td>1.5</td>
<td>3</td>
<td>0.65</td>
</tr>
<tr>
<td>0.66</td>
<td>2</td>
<td>0.75</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>0.75</td>
<td>1</td>
<td>0.38</td>
<td>1</td>
<td>0.21</td>
</tr>
<tr>
<td>0.82</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>0.21</td>
</tr>
<tr>
<td>1</td>
<td>242</td>
<td>90.98</td>
<td>425</td>
<td>91.2</td>
</tr>
<tr>
<td>Total</td>
<td>266</td>
<td>100</td>
<td>466</td>
<td>100</td>
</tr>
</tbody>
</table>

Table 7: Distribution of the family validation according to the precision rate

<table>
<thead>
<tr>
<th>Precision</th>
<th>Number of families</th>
<th>Percentage</th>
<th>Number of families</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>169</td>
<td>63.53</td>
<td>324</td>
<td>69.52</td>
</tr>
<tr>
<td>0.12</td>
<td>1</td>
<td>0.38</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>0.16</td>
<td>2</td>
<td>0.75</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>0.25</td>
<td>3</td>
<td>1.12</td>
<td>1</td>
<td>0.22</td>
</tr>
<tr>
<td>0.285</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>0.22</td>
</tr>
<tr>
<td>0.333</td>
<td>5</td>
<td>1.88</td>
<td>11</td>
<td>2.36</td>
</tr>
<tr>
<td>0.42</td>
<td>1</td>
<td>0.38</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>0.5</td>
<td>9</td>
<td>3.38</td>
<td>17</td>
<td>3.64</td>
</tr>
<tr>
<td>0.6</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>0.43</td>
</tr>
<tr>
<td>0.66</td>
<td>1</td>
<td>0.38</td>
<td>5</td>
<td>1.07</td>
</tr>
<tr>
<td>0.71</td>
<td>1</td>
<td>0.38</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>0.75</td>
<td>1</td>
<td>0.38</td>
<td>2</td>
<td>0.43</td>
</tr>
<tr>
<td>0.875</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>0.22</td>
</tr>
<tr>
<td>0.909</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>0.22</td>
</tr>
<tr>
<td>1</td>
<td>73</td>
<td>27.44</td>
<td>101</td>
<td>21.67</td>
</tr>
<tr>
<td>Total</td>
<td>266</td>
<td>100</td>
<td>466</td>
<td>100</td>
</tr>
</tbody>
</table>

3.3.3. Limits of the evaluation with recall and precision rates
The validation of the results shows that the rules infer not only synonymy links, but also other types of semantic relations such as hyperonymy and see-also
Detection of synonymy links between terms: experiment and results

Although the precision is low (33% of the links have been validated), results are interesting for the terminology structuring. Indeed, while the acquisition of links by an expert would take tens of hours, the validation of the links automatically inferred on the DSE corpus took three hours.

The recall rate seems very moderate. It is difficult to evaluate. There is no golden standard which would give the relevant relations of the document. The comparison with links manually detected by an expert is biased as manual detection is subjective. It is well-known that various experts give different results depending on their various backgrounds (Szpakowicz et al. 1996). Moreover, building a consensual terminology from several ones is a real problem because the terminology structuring itself varies from one terminologist to another (Roulin and Cooper 1993, Srinivasan 1992).

However, from the terminologist's point of view, recall rate must be increased, even at the precision's expense. As opposed to information retrieval systems, the precision rate is less important than the recall in terminological tools (Jacquemin 1994). One way to increase recall is to combine various resources. The collaboration with EDF allows to evaluate the complementarity of different resources. Specialized lexical resources are available for the DSE corpora. We report results of the semantic links detection using these various sources (section ).

On the other hand, we note that the precision rate does not reflect the help to the terminology building. Results point out unobvious semantic relations between the terms of a document and suggest new links. Semantic links are generally different from those found by the experts: terminologists rather focus on specialized terms already attested. This contrast between the helpfulness of our algorithmic approach and the interpretation of the precision rate raises the problem of the evaluation of our tools. We propose a new measure to evaluate results in the context of a terminology structuring tool (section ).

4. Which lexical source for which corpus?

4.1. Using lexical sources with technical corpora

Endogenous approaches are insufficient for terminology building. The acquisition of specialized knowledge from technical documents requires methods which combine existing knowledge sources and specialized corpora. Various lexical sources can be used.
Resources are generally specific for the domain of the working corpus. In the context of a document filtering tool, Naulleau (1998) builds semantic classes starting with information extracted from the EDF Thesaurus. Acquisition of new hyperonymic links from an agronomy corpus is bootstrapped with relations provided by a thesaurus of the domain (Morin 1998). Habert et al. (1998) aim at extending and adapting a medical nomenclature with the Menelas corpus which deals with coronary diseases. Term disambiguation method proposed by Maynard and Ananiadou (1998) is based on a corpus word distribution and on UMLS thesaurus.

As general semantic information has often been considered as irrelevant for specialized language, it has rarely been used for knowledge acquisition from technical corpora. However, Basili et al. (1997) and Basili et al. (1998) show how WordNet and LDOCE can be tuned and specialized for technical corpora. We adopted a similar approach for terminology structuring by combining a candidate term list extracted from a corpus and the synonymy links of a general language dictionary (Hamon et al. 1998).

However, few works have studied the respective contribution of these two types of resources (general and specialized) for terminology building. The reported study tries to draw light on their respective usefulness.

In the context of our collaboration with EDF, we tried to identify the best way to bootstrap our semantic links acquisition. We therefore tested our approach with various lexical sources from the most to the least general ones: Le Robert dictionary, a technical thesaurus, and some very specific hand built semantic classes. The following sections describe these experiments.

4.2. Using manually built semantic class as a bootstrap

For the DSE2 corpus, synonymy classes manually built by a terminologist are available. Terms extracted from a different corpus of the same domain have been studied. The terminologist regrouped 1 335 terms in 500 classes. We have considered those classes as equivalence classes. 3 456 synonymy links have been extracted. Those links are morpho-syntactic relations appoint en acide borique (supplement in boric acid) / appoint en bore (boron supplement) or semantic relations eau de refroidissement (cooling water) / fluide réfrigérant (cooling fluid). Contrary to Le Robert's synonyms, links may hold between two complex terms or between a simple and a complex terms appareil de mesure (measuring device) / capteur (sensor). Such specialized resources are important in
Detection of synonymy links between terms: experiment and result

Terminology structuring. Unfortunately, they are rare and time-consuming to build.

We applied the inference rules on the DSE$_2$ corpus starting with the links provided by the semantic classes. Among the 3,456 links, only 281 links are used by the rules to infer 169 links between complex terms. The expert validated 115 of the links (68%).

We also tried to combine these two different sources. This help to understand the respective contribution of each source type. Besides the 590 links inferred from the dictionary and the 169 links inferred from the classes, 44 new links are detected thanks to the third rule (which takes advantage of both lexical sources jointly). Indeed, links such as tronçon du circuit de réfrigération intermédiaire (section of the intermediate cooling circuit) / portion du circuit RRI (portion of the RRI circuit) need two initial links issued from the different sources. The dictionary provides the links tronçon (section) / portion (portion) while circuit de réfrigération intermédiaire (intermediate cooling circuit) / circuit RRI (RRI circuit) belongs to a semantic class. Among the links inferred form the dictionary and the semantic classes, 266 links (39%) have been validated by the expert.

4.3. Inferring links from a thesaurus

We also attempted to infer links from the EDF thesaurus (EDFDOC). This thesaurus is composed of 20,000 simple or complex terms. Terms are laid out into 330 semantic fields which are grouped into 45 points of view. These are general classes which can be viewed as the top of an ontology. Terms are linked with three types of relations: hierarchical one (hyponymy), associative one (see-also) and synonymy one (is used for). We exploited all the semantic links i.e. 25,000 links.

Although most of the thesaurus links are morpho-syntactic or semantic relations between complex terms, we attempted to saturate the terminological network as for the semantic classes. The rules used only 389 links of the thesaurus and inferred 55 links. The expert validated 36 of the links (65.45%). As thesaurus initial terms are already complex, only few links between more complex terms can be inferred.
4.4. Comparison of the results

4.4.1. Global comparison

Table 8 summarizes results obtained from the three sources. We define the initial links as the links resulting of the filtering step i.e. the links provided by a lexical resources for which both linked words are present in the corpus. Inferred links are defined as the links proposed by the inference rules.

The ratio of inferred links according to the links used by the rules varies from one source to another: 3/5 for the semantic classes and only 1/5 for the dictionary. Nevertheless, 78% of the links are inferred from the dictionary while only 22% for the semantic classes. This is due to the dictionary large coverage. It must be noted that using semantic information from the dictionary is far less time-consuming than the building and the use of semantic classes. The thesaurus stands in the worst position. Both its inferred links ratio (1/7) and its coverage are low. The distribution of the links according to the rules are identical whatever the lexical source: 62% of the links have been inferred with the rule 2. The rule 3 is the less productive.

Table 8: Results of the detection of semantic links

<table>
<thead>
<tr>
<th></th>
<th>Link filtering</th>
<th>Inference of links on the DSE corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of initial links</td>
<td>Number of terms</td>
</tr>
<tr>
<td>Dictionary</td>
<td>3 129</td>
<td>1 299</td>
</tr>
<tr>
<td>Semantic classes</td>
<td>281</td>
<td>344</td>
</tr>
<tr>
<td>EDF Thesaurus</td>
<td>389</td>
<td>478</td>
</tr>
<tr>
<td>Dictionary then classes</td>
<td>3 376</td>
<td>1 547</td>
</tr>
</tbody>
</table>

We characterized the ratio of morpho-syntactic links found with our approach. Three types of relation are considered: derivation, ellipsis and etymology. Only 28 validated links are morpho-syntactic: 16 links are derivational relations (*ligne des drains* (line of drain) / *ligne de drainage* (draining line)), 4 links are ellipsis relations (*dégazage du circuit TEP* (outgassing of the TEP circuit) / *dégazage du TEP* (outgassing of the TEP)) and 8 links are etymological relations (*refroidissement des pompes primaires* (cooling of the primary pumps) / *réfrigération des pompes primaires* (cooling of the primary pumps)). We compare these results with the links that FASTR’s analysis (Jacquemin96) propose (this
Detection of synonymy links between terms: experiment and results

tool is designed to detect morpho-syntactic variants of terms are available for our experiments: only 4 derivational links are common (classement sismique (seismic classification) / classification sismique (seismic classification)). Although no definitive answer can be given from the few examples, they suggest that our dictionary-based approach is complementary to methods which have been specifically design for the detection of morpho-syntactic variants. It proposes relations that tools as FASTR do not detect (ellipsis and etymological relations).

4.4.2. Complementarity of the semantic classes and the dictionary

The semantic classes are not only used as a lexical resource. They reflect a manual terminological work to which the results obtained through automatic detection can be compared. The Table 9 presents the results according to the number of inferred links present in the semantic classes i.e. the inferred links already built by the expert.

Table 9: Distribution of the links built by the expert among the inferred links

<table>
<thead>
<tr>
<th></th>
<th>Number of inferred links built by the expert</th>
<th>Number of links which are not built by the expert</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semantic classes</td>
<td>25</td>
<td>144</td>
</tr>
<tr>
<td>Dictionary</td>
<td>23</td>
<td>567</td>
</tr>
<tr>
<td>Dictionary then classes</td>
<td>38</td>
<td>720</td>
</tr>
</tbody>
</table>

One interesting fact is that new links can be inferred from the semantic classes: this shows that the expert did not saturate the classes. Only few links built by the terminologist have also been automatically inferred. For instance, link condition de fonctionnement (condition of running) / régime de fonctionnement (running activity) exists in a semantic class and is inferred from the link (condition condition / régime activity) given by the terminologist. 144 links are inferred form the semantic classes and only 25 out of 144 links are already present in the semantic classes. It seems that while building the classes, the terminologist focuses on their consistency rather than their inferential enclosure. This shows that the human and algorithmic approaches are complementary for the detection of synonymy links.

Moreover, few links inferred from the dictionary are given by the terminologist in the semantic classes (23 out of 590 links). The expert seems to focus on the technical language. For him, the detection of links between terms existing in the
general language is time-consuming. Nevertheless, during the validation, he considered those links as useful for a document consulting system.

4.4.3. The thesaurus: a weaker aid

Similarly, the overlap of links inferred from the dictionary or the semantic classes and the links provided by the thesaurus is low. Only two links are common to the links inferred from the dictionary and the thesaurus while one link is both inferred from the semantic classes and provided by the thesaurus (see Table 10).

Table 10: Distribution of the thesaurus links among the inferred links

<table>
<thead>
<tr>
<th></th>
<th>Number of links occurring in the thesaurus</th>
<th>Number of links not occurring in the thesaurus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dictionary</td>
<td>2</td>
<td>588</td>
</tr>
<tr>
<td>Semantic classes</td>
<td>1</td>
<td>168</td>
</tr>
<tr>
<td>Dictionary then classes</td>
<td>2</td>
<td>756</td>
</tr>
</tbody>
</table>

The domain of the corpus is partially covered by the thesaurus. Only 28 links are common to the thesaurus and the semantic classes which are built from a different corpus of the same domain.

It seems that the thesaurus stands in an intermediate position between the general language dictionary and specialized resources such as semantic classes. However, as mentioned above, it combines the weaknesses of both.

This comparison of the semantic links inferred from various lexical sources shows the usefulness of the general language dictionaries to process specialized corpora. When they exist and even if they have been built for similar corpora, specialized lexical resources remain incomplete and are often very technical. A general dictionary therefore represents a precious complementary knowledge source.

5. How to evaluate the results?

5.1. Evaluation problem in terminology

Evaluation of terminological tools raises difficulties. There is no golden standard to compare with and traditional measures of recall and precision are often
Detection of synonymy links between terms: experiment and results

irrelevant. It is also noticeable that most evaluations deal with term extraction rather than terminology structuring.

5.1.1. Terms extraction
A first approach to evaluate term extraction tools is global and quantitative. El-Hadi et Jouls (1998) argues that evaluation must take into account recall and precision rate according to a terminology reference. However, this approach assumes that such terminology is available and that the comparison of results of different indexing systems can be done.

One can oppose to that a qualitative evaluation (Sparck-Jones and Galliers 1996). Following L'Homme et al. (1996) and Bourigault and Habert (1998), we argue that one should first point out evaluation criteria and identify the relevance of the contributions and limits of each tool according to various tasks.

Actually, as in most NLP fields, the terminology evaluation widely depends on the task and the application (Resnik and Yarowsky 1997). Evaluation criteria (number of terms, type of semantic relations, accuracy of the description) depend on the terminology use: the terminologies required for document consulting system, text authoring, document classification or ontology building are different. For instance, links inferred from the synonymy link ligne (line) / tuyauterie (piping) may be validated in the context of a consulting system and prohibited in the context of text authoring tool. This point of view is all the more important, in the context of aiding tools, for which the user's control must be taken into account.

5.1.2. Terminology structuring
As regards the terminology structuring, there is almost no evaluation experiments. The problem has been pointed out (Mayfield and Nicholas 1992, Assadi 1997). Evaluation measures are partial. For instance, in the system PROMETHEE (Morin 1998) which acquires hyponymic relation, each lexicosyntactic extraction pattern is evaluated independently. This helps the parameter tuning but the system and the method are not evaluated as a whole.

According to the terminologist's point of view, the structuring tools are helpful. We showed on the DSE1 corpus that the automatically inferred links are different and complementary to the semantic classes which reflect the links manually added by the terminologist. The algorithmic approach suggests new links.
5.2. A new evaluation measure

It is important to account for the usefulness of the approach and the suggestiveness of the links. We propose a measure to evaluate the aid of our system for the terminologists in the context of a terminology structuring. This measure takes into account the characteristics of our system and evaluates it according to the structure of the results.

5.2.1. Minimization of the errors in the homogeneous families

The analysis of results showed that some initial links provided by the dictionary produce numerous erroneous links. These errors are quick to detect because they are similar and they can be eliminated all at once. The new evaluation measure minimizes the error weight in the homogeneous families.

The error ratio of the family $i$ is computed from the number of errors in $i$. We set this ratio $R_i$ as:

$$R_i = \frac{\max(NbLinks_i / 2, 1)}{NbErrors_i}$$

if $NbErrors_i > 0$

where $NbLinks_i$ is the size of the family and $NbErrors_i$ is the error number.

If the family is homogeneous and all the links have been rejected, this ratio is $\frac{1}{2}$. It increases if the numbers of errors decreases according to the size of the family. When the number of validated links and of invalidated links are equal, the ratio is 1. If the family contains a single link, the ratio is 1.

We define the weight $W_i$ of an error in a family $i$ as:

$$W_i = \min(R_i, 1)$$

It varies between $\frac{1}{2}$ and 1.

This leads to define a weighted precision:

$$\text{WeightedPrecision} = \sum_{i=1}^{N_f} \frac{NdErrors_i \times W_i}{NbLinks_i}$$

where $N_f$ is the number of families.

The factor $W_i$ is used to minimize the weight of each error in the families where most of the links are invalidated, for instance temps (time) / température (temperature). Compared with the classical precision, the weighted precision
Detection of synonymy links between terms: experiment and result

better reflects the terminologist’s point of view and the usefulness of the inferred links for the terminology structuring.

5.2.2. Evaluation of the results on the working corpora
We compute the weighted precision on two corpora. The Table 11 summarizes the results according to this measure on the DSE₁ corpus and the DSE₂ corpus. The weighted precision is 49.58% for the former corpus and 59.07% for the latter one while the classical precision is 37% on the DSE₁ corpus and 33.72% on the DSE₂ corpus.

Table 11: Comparison of the weighted and classical precision

<table>
<thead>
<tr>
<th></th>
<th>DSE₁ corpus</th>
<th>DSE₂ corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weighted precision</td>
<td>49.58%</td>
<td>37%</td>
</tr>
<tr>
<td>Classical precision</td>
<td>50.18%</td>
<td>33.72%</td>
</tr>
</tbody>
</table>

5.3. Discussion
It is difficult to appreciate the relevancy of such an evaluation measure. Other experimentation in various application contexts would be required to get a better idea. One difficulty is that formally evaluating the cognitive cost of the validation process for the expert would call for specific psychological and ergonomic tests.

However, it is well-known that the way the terminological results are presented affects the validation. This presentation varies with the system and the application. Thus, the results of the LEXTER corpus analysis are presented as a hypertextual syntactic network. Assadi (1997) also tries to take these ergonomic aspects into account. We propose two representations of the results: classes built by transitive enclosure and families. The above weighted precision is therefore a first attempt to take the structure and the presentation of the results into account for their evaluation.

The way the results are presented may also make the terminologist think of new links. This suggestion process is an important one. We noted in a first experiment that the expert considered half of the links as interesting even if he validated only 37% of them as real synonyms. Morin (1998) make similar observation for experiments with PROMETHEE. Contrary to him, we choose to retain these links as good ones.
Wrong links such as *groupe froid* (cold group) / *groupe frigorifique* (refrigerating group) and *groupe froid* (cold group) / *unité de refroidissement* (cooling unit) may also suggest a good one *groupe frigorifique* (refrigerating group) / *unité de refroidissement* (cooling unit). The expert underlined himself that correcting links is far less difficult than finding new links spontaneously. Unfortunately, up to now, this last property has not been properly evaluated. This remains to be done.

6. Conclusion
The collaboration with EDF lead us to experiment our synonymy detection system in a realistic context. By interacting with a terminologist who makes use of the resulting terminology, we got an expert feedback which helps us to understand the usefulness of our approach and the best way to handle it.

Globally the results on the DSE$_2$ corpus confirms our previous experimental observations. The traditional precision measure and the distribution of detected links among families of different size are close.

The present experiment helps to understand the contribution of a general language dictionary for technical text processing. We carried out experiments to compare the synonymy links inferred from *Le Robert* dictionary with the semantic information given by two specialized hand built lexical resources. It appears that general language dictionary represents a useful knowledge source which is complementary to specialized ones.

The industrial collaboration also enabled a better evaluation of the resulting structured terminology in the context of a document consulting task. The precision and recall measures are difficult to appreciate and do not appear as very relevant from the terminologist's point of view. We therefore proposed a weighted precision measure which better reflects the practical usefulness of the results.

However several evaluation problems remain to be solved. As in several terminological works, we have underlined that errors may nevertheless be suggestive and thus interesting for the terminologists. This property is difficult to take into account for evaluation. On the other hand, as there is no golden standard to which the results can be compared, it is difficult to measure the silence of terminological tools. Recall is however an important piece of information for evaluation. In order to reduce erroneous links, we plan to test
Detection of synonymy links between terms: experiment and results.

various relevance measures based on contextual distance and information extracted from the resources.

Acknowledgment

This work results from a collaboration with the Direction des Etudes et Recherche (DER) d'Electricité de France (EDF). We thank Henry Boccon-Gibod, Yasmina Abbas, Marie-Luce Picard (DER-EDF), Didier Bourigault (CNRS). The discussions we had with them as well as Christian Jacquemin (Limsi), Benoît Habert (UMR 8503), Christophe Fouqueré (LIPN) have contributed to this work.

References


Jacquemin, C. 1996. “A Symbolic and Surgical Acquisition of Terms Through Variation”. In Connectionist, Statistical and Symbolic Approaches to Learning for
Detection of synonymy links between terms: experiment and resul


Maynard, D. and Ananiadou, S. 1998. “Acquiring Contextual Information for Term Disambiguation”. In *Proceedings of COMPUTERM'98 (First Workshop on Computational Terminology) – Coling-ACL'98*, 86-90. Université de Montréal, Montréal, Quebec, Canada.


As the terms used in this paper have been extracted from French documents, their translation, especially for the synonymy, does not always express the same nuance than originally.

Coronarography manual, physician letters, Patient Discharge Summaries.

A connected component is a graph such from a node, one can access to any other.

The validation included not only the links inferred from the general language dictionary, but also links inferred from other resources (see below).

One can argue that consider etymology relation in the morpho-syntactic relation is not obvious. For a sake of simplicity, we group etymology with other morpho-syntactic relation.

ligne refers to the device and the tuyauterie to the object. According to the expert, the meaning of these words is not equivalent in the domain even if they are often used as if they were.

For instance, two following links form a class: échelle logarithmique (logarithmic scale) / gamme logarithmique (logarithmic range) and échelle logarithmique (logarithmic scale) / mesure logarithmique (logarithmic measure).