Bring vs. MTRoget: Evaluating Automatic Thesaurus Translation

Lars Borin,* Jens Allwood,† Gerard de Melo‡

*Språkbanken, Dept. of Swedish
University of Gothenburg, Sweden
lars.borin@svenska.gu.se

†SSKKII, Dept. of Applied IT
University of Gothenburg, Sweden
jens.allwood@gu.se

‡IIIS, Tsinghua University
Beijing, P.R. China
demelo@tsinghua.edu.cn

Abstract

Evaluation of automatic language-independent methods for language technology resource creation is difficult, and confounded by a largely unknown quantity, viz. to what extent typological differences among languages are significant for results achieved for one language or language pair to be applicable across languages generally. In the work presented here, as a simplifying assumption, language-independence is taken as axiomatic within certain specified bounds. We evaluate the automatic translation of Roget's *Thesaurus* from English into Swedish using an independently compiled Roget-style Swedish thesaurus, S.C. Bring's *Swedish vocabulary arranged into conceptual classes* (1930). Our expectation is that this explicit evaluation of one of the thesaureses created in the MTRoget project will provide a good estimate of the quality of the other thesauruses created using similar methods.

Keywords: lexical resource, Roget, Swedish

1. Introduction

Evaluation of automatic language-independent methods for language technology resource creation is difficult, since the situations where you will need to resort to automatic methods almost by definition are such that there will be no gold-standard evaluation data available. The applicability of such methods is also confounded by a largely unknown quantity, viz. to what extent typological differences among languages are significant for results achieved for one language or language pair to be applicable across languages generally (Bender, 2011).

However, under certain additional assumptions, language-independence can be taken as axiomatic within specified bounds, and the case described here is arguably one where this holds.

Outside the NLP field, the most well-known lexical-semantic resource for English is without doubt Roget's *Thesaurus*[1] (Roget, 1852; Hüllen, 2004). Although not as well-known in NLP as the Princeton WordNet (Fellbaum, 1998), the digital version of Roget offers a valuable complement to WordNet (Jarmasz and Szpakowicz, 2004), which has seen a fair amount of use in NLP (e.g., Morris and Hirst 1991; Jobbins and Evett 1995; Jobbins and Evett 1998; Wilks 1998; Kennedy and Szpakowicz 2008).

Similarly to WordNet, it would be desirable to have counterparts to Roget for other languages than English. While wordnets have been built for or translated into a number of other languages,[2] translations of Roget are rare. For this reason, de Melo and Weikum (2008b) proposed a method for translating Roget to French fully automatically. Their proposed method can be abstractly characterized as belonging to a class of methods involving transfer of annotations from one language to another via bilingual resources, using word and phrase translations as the bridging elements (see section 2).

The present paper is devoted to the problem of evaluating such automatically created resources, capitalizing on the existence of a human-authored Swedish Roget-style thesaurus (see section 3). Thus, we present an explicit evaluation of MTRoget-swe, an automatic translation of Roget to Swedish, by way of explicit comparison with this resource, following the advice of Sahlgren and Karlgren (2005), who argue that a lexicon is best evaluated by comparing it with an existing lexicon of the same type.

Our expectation is that the results of the explicit and very detailed quantitative and qualitative comparison described below will carry over to other languages – at least under specific additional assumptions to be discussed below – and more generally shed light on the merits and challenges of automatic translation approaches to building lexical resources.

2. Automatic Thesaurus Translation

Roget's *Thesaurus*, first published by Peter Mark Roget in 1852, ranks among the most well-known reference works on the English lexicon and is based on a conceptual classification of words into slightly over 1,000 hierarchically organized classes (Old, 2004).[3] In each class, there can be separate lists of relevant nouns, verbs, adjectives, adverbs, and phrases.

Semicolons, together with paragraph structure, are used to organize words into smaller groupings, which are thought to be more closely semantically related. The American 1911 edition (Roget, 1911) has been made available as a text file through Project Gutenberg by Cassidy (2000), with minor extensions, including more than 1,000 new terms and annotations that mark obsolete and archaic forms.

[1] Also alternately referred to as "Roget" below.

[2] See the *Global WordNet Association* website: <http://globalwordnet.org>.

[3] In this paper, we use the term "class" in reference to the numbered sections in Roget, which are often also called "head(word)s". But note that Roget's Thesaurus itself reserves the word "CLASS" for the highest-level subdivisions (e.g. CLASS I: "WORDS EXPRESSING ABSTRACT RELATIONS"), of which there are only very few.

Our MTRoget project is based on the idea of automatically translating the English thesaurus using machine-readable dictionaries. Although machine-readable in a trivial sense, the Gutenberg version by Cassidy (2000) is simply a digital plaintext rendering of Roget (1911), which of course was designed exclusively for human readers.

Parsing this text file in order to obtain a hierarchy of headings and classes is not quite as trivial as it may seem at first sight. Not only does a myriad of implicit formatting and structuring conventions need to be accounted for, but also the fact that the source file frequently fails to abide by the inferred syntax rules defining well-formed entries, as there are a considerable number of inconsistencies and formatting errors. We used a recursive top-down approach to identify the six top-level groupings, which include e.g. "words relating to the intellectual faculties", and then proceed to deeper levels. The top-level groupings are sometimes subdivided into divisions, e.g. "communication of ideas", which consist of sections, e.g. "modes of communication".

Sections can be further subdivided into multiple levels of subsections, which finally contain classes. Under each class one finds one or more part-of-speech markers followed by groups of terms or phrases relating to the class. These groups are delimited by semicolons or full stops, and within such groups, commas or exclamation marks usually fulfil the function of separating individual items, though care needs to be taken not to split up phrases containing such characters. In addition to terms and phrases, these 'semicolon groups' may also contain references to other classes or to other parts-of-speech of the current class (see figure 1).

We have consequently developed a fairly complex parsing process to bring the semi-structured information in the Roget's Thesaurus text file into a structured form, in our case a list of subject-predicate-object triples that describe word membership in classes (as well as a few other semantic relationships).

Our approach to automatical translation of Roget to a number of languages is based on previous work for producing a French translation of Roget's Thesaurus (de Melo and Weikum, 2008b).

There, we designed a disambiguation approach on the basis of a technique which we had initially developed to generate a German-language version of WordNet (de Melo and Weikum, 2008a) that has now been extended to include several novel statistics. The basic idea is to use supervised machine learning to derive a model for classifying translations from manually labelled translation pairs. We conceive each semicolon group in the thesaurus as a separate node to be translated to one or more terms.

Since a good coverage of the target language is an important desideratum, we allowed for translating a single English term to multiple French terms whenever this is appropriate. Furthermore, nodes may also remain vacuous when no adequate translation is available, as many thesauri are designed to cover a wide range of terms, including rare and obsolete terms that may often be untranslatable. This is most certainly the case for Roget's Thesaurus, bearing in mind that merely 41% of the terms in the 1987 Penguin Edition are covered by WordNet 1.6 according to Jarmasz and Szpakowicz (2001).

Given a French target term t and a thesaurus node n, we considered the tuple (n, t) a candidate mapping if and only if one of the English source terms associated with n in the original thesaurus is translated as t according to the unified translation knowledge base. Such tuples can either represent appropriate translations (used as positive training examples) or inappropriate ones (used as negative training examples).

Generalizing this approach to other target languages, an English word e in a given thesaurus class represented by node n can then be translated with zero, one, or more non-English terms t in any other language. For each new target language, suitable bilingual lexical resources are required. Thus, to build MTRoget-swe, a Swedish version of Roget's Thesaurus, such word translations were extracted from a number of freely available sources on the Web:

- the English Wiktionary
- the Swedish Wiktionary
- Apertium
- FreeDict
- GEMET
- OmegaWiki
- Magic-Dic

Acceptable translations t for a given thesaurus class n are distinguished from unacceptable ones by computing a range of features between n and t and then invoking a supervised discriminator, trained on manually annotated (n, t) pairs.

For example, scores like

$$\sum_{e\in\phi(t)} \frac{m}{m + \sum\limits_{n'\in\sigma(e)} (1 - \mathrm{sim_n}(n, n'))}$$

with $m = \max\limits_{n'\in\sigma(e)} \mathrm{sim_n}(n, n')$

characterize how many dissimilar alternatives n' there are to n in the set of classes $\sigma(e)$ for each English word e in the set of translations $\phi(t)$. Here, $\mathrm{sim_n}$ is a graph-based similarity measure between thesaurus nodes. After computing these scores for the training data pairs, we are able to use them as features to train a model that predicts whether new Swedish words t should be assigned to particular classes n.

Rather than setting the translation classes to correspond to the very coarse-grained 1,000-odd Roget classes, we in fact considered groups of words separated by semicolons or full stops as the target nodes n, as such a fine-grained classification has proven useful in a range of NLP tasks in the past.

The training data consisted of 731 manually labelled candidate (n, t) pairs. The words t in the training set were French language ones. However, as the features of the model are essentially graph and similarity-based, the learned model has

```
137. Infrequency -- N. infrequency, rareness, rarity; fewness &c 103;
seldomness^; uncommonness.
V. be rare &c adj..
Adj. unfrequent^, infrequent; rare, rare as a blue diamond; few &c 103;
scarce; almost unheard of, unprecedented, which has not occurred within
the memory of the oldest inhabitant, not within one’s previous
experience; not since Adam^.
     scarce as hen’s teeth; one in a million; few and far between.
Adv. seldom, rarely, scarcely, hardly; not often, not much,
infrequently, unfrequently^, unoften^; scarcely, scarcely ever, hardly
ever; once in a blue moon.
     once; once in a blue moon; once in a million years; once for all,
once in a way; pro hac vice [Lat.].
Phr. ein mal kein mal [G.].
```

Figure 1: Excerpt from the Roget’s Thesaurus text file: class/head(word) 137 (*Infrequency*)

```
137. ovanlighet; sällsynthet, sällspordhet, undantag,
undantagsförhållande, undantagsfall, undantagsställning, raritet,
rarhet, våplycka, vit korp, tunnsåddhet, fåtalighet, knapphet, brist.

v. tryta, fattas, brista.

a. ovanlig, sällsynt, sällspord, undantagsmässig, rar, enastående,
enstaka, gles, tunnsådd, fåtalig, sparsam, knapp;
   oerhörd, ohörd, utomordentlig, exempellös, häpnadsväckande, fenomenal;
   sällan, undantagsvis, knappt, knappast någonsin, icke allom givet,
icke i mannaminne, någon enda gång, en och annan gång, ibland,
stundom, emellanåt, alltemellanåt, då och då, ömsom, sparsamt, här och
där, här och var, ont om, knappast, näppeligen.
```

Figure 2: Excerpt from the Bring text file: class/head(word) 137 (*Ovanlighet* ‘infrequency’)

been found to be applicable to terms *t* in Swedish and other languages as well. Using this model, we obtain MTRoget-swe, an automatically produced Swedish thesaurus that closely follows the structure of Roget’s Thesaurus (see figure 4, top).[4].

3. Bring’s Swedish Thesaurus

The author of what is probably the first Swedish thesaurus, Sven Casper Bring (1842–1931) worked as a lawyer, district judge and translator. Besides practicing law, he published several translations from French, Italian and English to Swedish. His final work was an adaptation of Roget’s *Thesaurus* to Swedish, which appeared in 1930 under the title *Svenskt Ordförråd ordnat i begreppsklasser* ‘Swedish vocabulary arranged in conceptual classes’. He writes in his preface to the book that he was inspired by similar adaptations that had taken place of Roget’s *Thesaurus* to German.

Like in Roget’s Thesaurus, the vocabulary included in the book is divided into slightly over 1,000 “conceptual classes”. A “conceptual class” corresponds to a class, or “head(word)”, in Roget’s Thesaurus. Each conceptual class consists of a list of words, where, when there are enough relevant words, nouns are listed first, followed by verbs, adjectives and last phrases. In some cases, Bring has not found words in all four categories. For instance, the class *libertin* ‘libertine’ contains only nouns, although this may not be Bring’s doing, since the same is true of the corresponding class in Roget, as well.

Following the structure introduced by Roget, semicolons, together with paragraph structure, group words together, which are thought to be more closely semantically related. In this way, the number of semicolon enclosed paragraphs to some extent indicate how richly structured a particular conceptual area is. Within each conceptual class and within each paragraph group, there is an attempt to list those words first, which are most similar to the word that identifies the “conceptual class” (see figure 2).

In the early 1990s, after having used Bring (1930) for several years in doing various kinds of semantic analysis, Jens Allwood secured funding allowing the digitization of Bring. The first digital version was ready in 1997, and

[4] MTRoget-swe is freely available from <http://www.demelo.org/gdm/mtroget/>

subsequently Grönqvist (2005) constructed a computerized browser for Bring.

The work on the digital version of Bring stopped temporarily after this, but in 2011 Språkbanken[5] agreed to make Bring available and maintain it as a component of its integrated lexical macroresource for Swedish language technology (Borin et al., 2010; Borin et al., 2012; Borin et al., 2013b).

As a result of this ongoing work, Bring has been partially mapped to other lexical resources through SALDO, a full-scale semantic and morphological lexicon for modern Swedish (Borin et al., 2013a). In the process, a large number of remaining OCR errors have been corrected and many thousands of Bring entries have been marked as obsolete and not mappable to SALDO.

The plan is to finish the mapping and then use SALDO and other modern lexical resources available in Språkbanken in order to semi-automatically add modern vocabulary to Bring, and to publish both the original and the modernized version under an open-source license (CC-BY), and to integrate them into the lexical macroresource of Språkbanken.

4. Thesaurus Evaluation

Given the unique opportunity of two resources with similar structure being available, we proceed with a detailed comparison of the automatically produced MTRoget-swe (*M* for short) with the human-created and subsequently digitized Bring thesaurus (*B* for short).

4.1. Quantitative evaluation

The sizes of the involved datasets are shown in table 1.

	Roget	*MTRoget-swe*	*Bring*
entries	98,774	72,816	148,606
unique lemma/POS	62,859	22,024	57,968

Table 1: The sizes of the lexical resources

Since the number of classes is not identical between Roget/MTRoget-swe and Bring – Roget and MTRoget-swe have 1,043, while Bring has 1,015 classes – we manually prepared a class mapping between the two thesauri. This mapping has 1,010 classes. In other words, Roget contains 33 classes missing from Bring (e.g., *nonuniformity*, *punctuality*, *sponge*, and *orthodoxy*), but on the other hand Bring adds five classes which are not found in Roget: *dag* 'day', *natt* 'night', *kropp* 'body', *väg* 'way, road, path', and *uppvaknande* 'awakening').

Using these data, we calculated how well MTRoget-swe covers Bring (see table 2).[6]

M ∩ B	M only	B only	M tot.	B tot.	B%	M%
21,145	41,240	126,660	62,385	147,805	14	34

Table 2: Coverage of Bring by MTRoget-swe

In table 2, **M tot** is the size in entries of the automatically translated resource, MTRoget-swe (in the 1,010 classes). **B tot** gives the number of entries in Bring in the 1,010 classes. Entries may be duplicates, i.e., the same lemma–POS combination may appear in more than one place. **M ∩ B** is the overlap, i.e., how many Bring entries in their correct class were generated by the automatic translation of Roget. **B only** and **M only** indicate the number of entries unique to MTRoget-swe and Bring, respectively.

If we calculate the number of overlapping entries in relation to the "target" – Bring – we get the percentage **B%**. Again, we only count entries falling into the correct Bring class. Comparing the overlap in entries to the result of the translation – MTRoget-swe – instead, we get the percentage **M%**.[7]

Instead of these aggregated figures, we can also investigate the coverage per Roget/Bring class (i.e., 1,010 classes that the two resources have in common). The coverage per class exhibits a highly non-uniform behavior. In figure 3, the 1,010 classes are ordered along the x axis according to their conventional numbering in Roget/Bring. The y axis indicates coverage in percent.

In figure 3, the upper jagged (red) curve shows the coverage in percent of MTRoget-swe entries per class (**M%** in the table above), i.e., which percentage of the automatically translated entries in a particular Roget class are also present in the corresponding Bring class. For a very few classes this number reaches 100%, and it never dips to zero.

The lower jagged (green) curve shows the coverage in percent of Bring entries per class (**B%** in the table above), i.e. how well the automatically translated Roget class covers the corresponding Bring class. In this case, we sometimes get no overlap at all, i.e., in some clasess, the automatic translation has not yielded a single corresponding Bring item.

In addition to the jagged curves, figure 3 also shows Bezier-smoothed coverage curves – which illustrate that **M%** and **B%** are not parallel, i.e., that the translation results are not uniform over the classes. The figure also shows (straight) regression lines for **M%** and **B%**.

Figure 4 shows our previous example class (137) in MTRoget-swe and the corresponding class in Bring (the latter repeated for convenience from figure 2). Common items are shown in uppercase. The coverage figures for this class are shown in table 3.[8]

[5] The Swedish Language Bank <http://spraakbanken.gu.se>

[6] In this comparison, the **M ∩ B** items (lemma/POS combinations) have to be in the same class. The differences in total number of entries compared to the figures given in table 1 are because (1) the comparison comprised only 1,010 classes, and (2) duplicate lemma/POS entries in the same class were eliminated before the comparison.

[7] In more familiar terms, we can think of **B%** as *precision* and **M%** as *recall*.

[8] Note that there are several duplicate lemma/POS-combinations in the MTRoget-swe class in figure 4. This is not surprising, given that word-level translation relations between any two languages are generally many-to-many. As mentioned previously, however, such duplicates have been ignored in all our calculations.

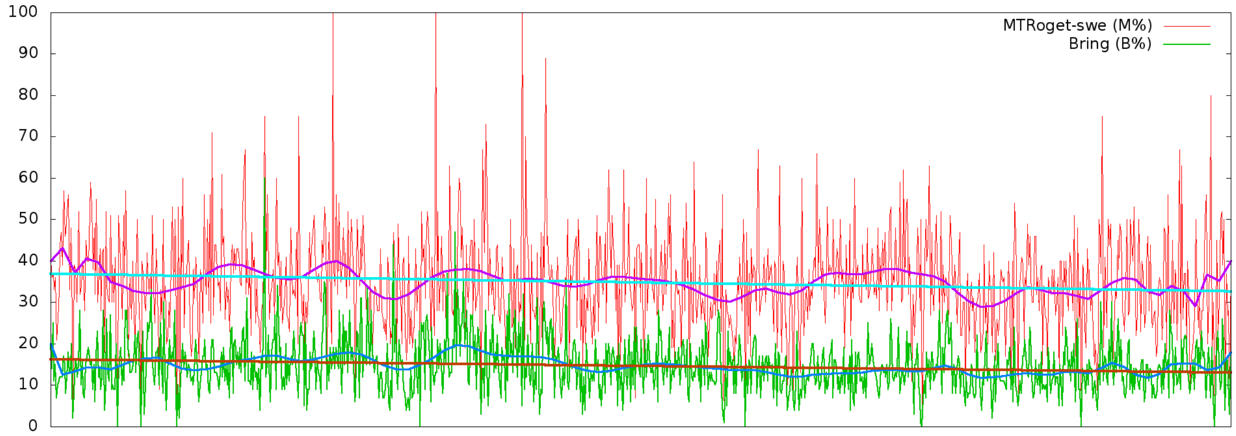

Figure 3: MTRoget-swe coverage of Bring (**M%** and **B%**) for the 1,010 classes

M ∩ B	M only	B only	M tot.	B tot.	B%	M%
12	5	44	17	56	21	71

Table 3: Coverage figures for class 137

We further calculated the overall MTRoget-swe vocabulary coverage of Bring when disregarding the class structure (see table 4). In other words, the figures in table 4 compare the entire vocabulary of MTRoget-swe with the entire vocabulary of Bring, "vocabulary" being defined as 'set of lemma/POS-combinations'. In this calculation all classes were used, not only the 1,010 common classes.

M ∩ B	M only	B only	M tot.	B tot.	B%	M%
13,753	8,271	44,215	22,024	57,968	23	62

Table 4: Coverage disregarding class structure

The low average coverage per class should be seen in light of the figures in table 4. We would have expected to find a larger share of the MTRoget-swe vocabulary in the much larger vocabulary of Bring. One reason for the discrepancy could be in the fact that Bring is an old dictionary, while the bilingual lexicons used for the translation of MTRoget-swe are modern. Since there is a manually prepared (partial) list of obsolete words in Bring, we performed a simple comparison of the MTRoget-swe vocabulary with this list (see table 5).

M ∩ B	M only	B only	M tot.	B tot.	B%	M%
113	21,911	5,740	22,024	5,853	2	0

Table 5: Coverage of obsolete entries in Bring

There is a marked difference; in table 5, only 2% of the obsolete Bring words are also in MTRoget-swe (against 23% for Bring as a whole). Compensating for these known obsolete words and estimating the final number of obsolete words,[9] we arrive at an adjusted Bring vocabulary coverage of approximately 33%.[10]

4.2. Qualitative evaluation

In the qualitative evaluation, we have systematically investigated a small random sample of the **M only** case. The sample comprised 200 items selected at random from the 41,240 **M only** entries. They could be classified into the four categories listed in table 6.

missing	different class	wrong POS same class	wrong POS different class
63	122	6	9

Table 6: Qualitative evaluation results

Looking more closely at the 63 missing items in table 6, we can classify them as shown in table 7.

modern words	missing words	wrong form	phrases	names	translation errors
21	12	13	10	2	5

Table 7: Breakdown of missing items

It is clear that a large share of the missing items are either modern words which entered the language after Bring was compiled, or words which are older (according to a standard historical dictionary of Swedish) but which for some reason do not appear in Bring. Some examples are given in table 8.

Somewhat surprisingly, we made the observation – encouraging in this context – that clear translation errors account for only a small part of the missing items. Some cases of wrong POS appear to have resulted from translation errors,

[9] About 14,000 Bring items which were not found in a large modern lexical resource remain to be investigated. reveals the majority of them to be obsolete, too.

[10] I.e., the result of dividing the overlapping 13,753 entries from table 4 by an estimated 40,000 non-obsolete Bring vocabulary items (out of a total of 57,968 items).

```
    #137. N. RARHET, RARITET, SÄLLSYNTHET; OVANLIGHET, SÄLLSYNTHET;
    Adj. SÄLLSYNT, RAR, OVANLIG; få; KNAPP, SÄLLSYNT, EXEMPELLÖS, utan
motstycke;
    Adv. KNAPPT, KNAPPAST, ovanligen, SÄLLAN, svårligen; KNAPPT, KNAPPAST,
svårligen; KNAPPT, KNAPPAST, svårligen; en gång;
```

```
137. OVANLIGHET; SÄLLSYNTHET, sällspordhet, undantag,
undantagsförhållande, undantagsfall, undantagsställning, RARITET,
RARHET, våplycka, vit korp, tunnsåddhet, fåtalighet, knapphet, brist.

v. tryta, fattas, brista.

a. OVANLIG, SÄLLSYNT, sällspord, undantagsmässig, RAR, enastående,
enstaka, gles, tunnsådd, fåtalig, sparsam, KNAPP;
   oerhörd, ohörd, utomordentlig, EXEMPELLÖS, häpnadsväckande, fenomenal;
   SÄLLAN, undantagsvis, KNAPPT, knappast någonsin, icke allom givet,
icke i mannaminne, någon enda gång, en och annan gång, ibland,
stundom, emellanåt, alltemellanåt, då och då, ömsom, sparsamt, här och
där, här och var, ont om, KNAPPAST, näppeligen.
```

Figure 4: Class 137 in MTRoget-swe (top) and Bring (bottom) with common items in uppercase

Item ‘gloss’	Description
aura/n ‘aura’	modern
haja/v ‘get the drift’	modern (colloquial)
senarelägga/v ‘postpone’	modern
ta av/v ‘take off’	modern orthography (Bring: *avtaga/v*)
vänsterjävel/n ‘left-wing bastard’	modern (colloquial)

Table 8: Missing items: modern words

Item ‘gloss’	Description: correct
besvära/v ‘bother’	wrong class
bord/n ‘table (furniture)’	wrong class
exotiska/n ‘exotic’	wrong lemma/POS: *exotisk/a*
lydigt/a ‘obedient(ly)’	wrong lemma: *lydig/a*
teleskop/v ‘telescope’	wrong POS: *teleskop/n*

Table 9: Missing items: mistranslations/misclassifications

and a part of the items ending up in the wrong class as well may have done so due to wrong translations given by the bilingual sources. See table 9.

While our comparison focuses on Swedish, we expect that under certain conditions many of the results carry over to other languages. Specifically, these conditions are (1) that comparable bilingual resources (dictionaries, parallel corpora, etc. of similar quality and coverage levels) are used for those language pairs, and arguably also (2) that the other languages are typologically similar (cf. Bender 2011, mainly with respect to the morphological complexity of words, which determines, directly or indirectly, the degree of accuracy we may expect, e.g., from automatic parallel corpus alignment.

Note, however, that the method evaluated here rests on the notion of *translation equivalence*, relying as it does on bilingual dictionaries, parallel corpora, etc., which actually presuppose that something of this sort exists. To the extent that we can find true translations of the terms in Roget, the construction of a corresponding lexical-semantic resource for a target language becomes an automatic affair, since the kind of lexical-semantic relations – (near-)synonymy and and (general) semantic closeness – which define the structure of Roget’s Thesaurus are assumed to be language-independent.

5. Conclusions and Future Work

Our evaluation showed the MTRoget-swe coverage of Bring to be fairly low, especially in the stricter sense of average coverage per class (34 **M%**/14 **B%**), but also in the looser sense of vocabulary coverage (62 **M%**/23 **B%**).

Returning to our original hypothesis, this would imply that MTRoget translations for other languages will yield similar coverage. However, a closer investigation revealed that the coverage figures can in part be explained by the fact that Bring and the translation dictionaries reflect different historical language stages. This complicates cross-language comparisons, since we cannot assume a uniform rate of vocabulary change across languages, and therefore cannot compute what the “true” expected coverage should be. Here, further research is needed.

In particular, we intend to contrast the coverage of Bring with the coverage of the bilingual resources that serve as inputs for the MTRoget approach, as this would aid in distinguishing genuine missed opportunities of the approach per se from more haphazard shortcomings of the input data, as indicated by terms present in Bring but missing from the bilingual resources.

Additionally, since a modernized Bring is in the works (see section 3), a natural follow-up study to that reported here is to evaluate MTRoget-swe against the modernized Bring.

Finally, using suitable bilingual dictionaries, the result of this evaluation can serve as new training data to bootstrap or filter additional lexicons for the MTRoget project.

6. References

Bender, Emily M. (2011). On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, 6(3).

Borin, Lars, Danélls, Dana, Forsberg, Markus, Kokkinakis, Dimitrios, and Toporowska Gronostaj, Maria. (2010). The past meets the present in Swedish FrameNet++. In *14th EURALEX International Congress*, pages 269–281, Leeuwarden. EURALEX.

Borin, Lars, Forsberg, Markus, Olsson, Leif-Jöran, and Uppström, Jonatan. (2012). The open lexical infrastructure of Språkbanken. In *Proceedings of LREC 2012*, pages 3598–3602, Istanbul. ELRA.

Borin, Lars, Forsberg, Markus, and Lönngren, Lennart. (2013a). SALDO: a touch of yin to WordNet's yang. *Language Resources and Evaluation*, 47(4):1191–1211.

Borin, Lars, Forsberg, Markus, and Lyngfelt, Benjamin. (2013b). Close encounters of the fifth kind: Some linguistic and computational aspects of the Swedish FrameNet++ project. *Veredas*, 17(1):28–43.

Bring, Sven Casper. (1930). *Svenskt ordförråd ordnat i begreppsklasser*. Hugo Gebers förlag, Stockholm.

Cassidy, Patrick. (2000). An investigation of the semantic relations in the Roget's Thesaurus: Preliminary results. In *Proceedings of CICLing 2000*, pages 181–204.

de Melo, Gerard and Weikum, Gerhard. (2008a). A machine learning approach to building aligned wordnets. In *Proceedings of the First International Conference on Global Interoperability for Language Resources*.

de Melo, Gerard and Weikum, Gerhard. (2008b). Mapping Roget's Thesaurus and WordNet to French. In *Proceedings of LREC 2008*, Marrakech. ELRA.

Fellbaum, Christiane, editor. (1998). *WordNet: An electronic lexical database*. MIT Press, Cambridge, Mass.

Grönqvist, Leif. (2005). *The Bring Browser – a computerized version of the Swedish Bring Thesaurus*. Department of Linguistics, University of Gothenburg.

Hüllen, Werner. (2004). *A history of Roget's Thesaurus: Origins, development, and design*. Oxford University Press, Oxford.

Jarmasz, Mario and Szpakowicz, Stan. (2001). The design and implementation of an electronic lexical knowledge base. In *Proceedings the 14th Biennial Conference of the Canadian Society for Computational Studies of Intelligence (AI 2001*, pages 325–333.

Jarmasz, Mario and Szpakowicz, Stan. (2004). Roget's Thesaurus and semantic similarity. In Nicolov, Nicolas, Bontcheva, Kalina, Angelova, Galia, and Mitkov, Ruslan, editors, *Recent Advances in Natural Language Processing III. Selected papers from RANLP 2003*, pages 111–120. John Benjamins, Amsterdam.

Jobbins, Amanda C. and Evett, Lindsay J. (1995). Automatic identification of cohesion in texts: Exploiting the lexical organization of Roget's Thesaurus. In *Proceedings of Rocling VIII*, pages 111–125, Taipei.

Jobbins, Amanda C. and Evett, Lindsay J. (1998). Text segmentation using reiteration and collocation. In *Proceedings of the 36th ACL and 17th COLING, Volume 1*, pages 614–618, Montreal. ACL.

Kennedy, Alistair and Szpakowicz, Stan. (2008). Evaluating Roget's thesauri. In *Proceedings of ACL-08: HLT*, pages 416–424, Columbus, Ohio. ACL.

Morris, Jane and Hirst, Graeme. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48.

Old, L. John. (2004). Unlocking the semantics of Roget's Thesaurus. In *2nd International Conference on Formal Concept Analysis*, pages 236–243, Berlin. Springer.

Roget, Mark Peter. (1852). *Thesaurus of English Words and Phrases*. Longman, London.

Roget, Mark Peter. (1911). *Thesaurus of English Words and Phrases*. Thomas Y. Crowell Company, New York.

Sahlgren, Magnus and Karlgren, Jussi. (2005). Automatic bilingual lexicon acquisition using random indexing of parallel corpora. *Natural Language Engineering*, 11(3):327–341.

Wilks, Yorick. (1998). Language processing and the thesaurus. In *Proceedings National language Research Institute*, Tokyo. Also appeared as Technical report CS–97–13, University of Sheffield, Department of Computer Science.