Compounds and other oddities in machine translation

Hanne Moa
Department of Language and Communication Studies
Norwegian University of Science and Technology
7491 Trondheim, Norway
hannemo@stud.ntnu.no

Abstract

ReCompounder is a working prototype for automatic translation of compounds by using the web as language model and a comprehensive bilingual dictionary as translation model. Evaluation shows this strategy is viable also with current search engine technologies.

In addition to a detailed overview of the system, there is a discussion of the current capabilities of the search engines used (Google and Yahoo! Search) and some tendencies relevant to lexicography. As the system currently translates from Norwegian to English, a brief introduction to compounding in Norwegian is included.

1 Introduction

A major problem in machine translation is what to do with words that are not found in the machine-readable lexicon.

A possible solution is to somehow connect a regular bilingual dictionary to the system, all the time keeping in mind that definitions are made in the lexicographical tradition and not finely tuned to the particular framework and methods used in the MT-system in question.

ReCompounder extends on the connected-dictionary idea. It is a proof-of-concept prototype which tries to translate previously unseen compounds. It is part of the LOGON machine translation project (Lønning et al., 2004) which aims to translate tourism[1]-relevant texts from Norwegian to English, using an approach based on semantic transfer.

The underlying idea is similar to what is done in Rackow et al.(1992) and Tanaka and Baldwin (2003), but using the world wide web as corpus. Grefenstette (1999) used a similar technique for example-based machine translation with the then popular search-engines.

Using a web search engine directly greatly limits what sort of translation is possible, but since the web and the indexes of the search engines continuously grow by the addition of new and newly written pages, the data one *can* analyze more closely resembles language as it used *just now*.

As for the rest of this paper, section 2 gives a brief overview of compounds in Norwegian and what they might translate to in English, section 3 presents ReCompounder in some detail, section 4 summarizes the first round of evaluation of the system, section 5 gives an overview of the results so far, and thereafter follows the conclusion.

2 The problem

Compounding as a word-formation process is *very* productive in Norwegian (Johannesen and Hauglin, 1996). Thus any MT-system translating to or from Norwegian needs a way to handle newly minted compounds that was not found in the system's lexicon.

The LOGON-project has access to several dictionaries and wordlists. Those which were used for this system and are referred to in this paper includes the Engelsk Stor Ordbok (Eek and others, 2001), Bokmålsordboka (Wangensteen, 2005) and the NorKompLeks-lists (Nordgård, 1996).

[1] More specifically hiking trip information

S. Werner (ed.), *Proceedings of the 15th NODALIDA conference, Joensuu 2005*, Ling@JoY 1, 2006, pp. 124–132
ISBN 952-458-771-8, ISSN 1796-1114. http://ling.joensuu.fi/lingjoy/

The Engelsk Stor Ordbok is a paper dictionary containing over 42000 Norwegian nouns translated to English. Bokmålsordboka is a dictionary of Norwegian while the NorKompLeks-lists are lists of Norwegian words in both inflected and uninflected forms and with pronunciations. The NorKompLeks lists were designed from the beginning to be machine-readable while the two others originally only existed on paper.

2.1 Compounding in Norwegian

The list of examples marked 1a) to e) below starts with a compound that were not found in any of the available dictionaries and wordlists, both text-versions and machine-readable, and derives other compounds from it:

(1) a. gårdshus
gård+hus
farmhouse

b. gårdshustak
gård+hus+tak
farmhouse roof

c. gårdshustakstein
gård+hus+tak+stein
farmhouse roof tile

d. gårdshustaksteinsproblem
gård+hus+tak+stein+problem
problem with the roof tiles of farmhouses

e. storgårdshus
stor+gård+hus
house of a large farm

This paper will not debate the definition of compounds, especially when or if a word form ceases to be a compound and instead turns into a proper lexical item. Instead a very pragmatic stance is taken: when it is *useful* to treat something as if it was a compound, it is treated as if it *is* a compound.

Most Norwegian compounds consists of noun-stems, as in example 1a) to d) above. Some nouns have compound-only suppletive stems, as in example 5 below.

The stems are in general *not* to be written with spaces, but occasionally they are separated by an epenthetic *s* or *e*. Compounds containing abbreviations are written with a hyphen, and newer compounds often are as well. See Johannesen and Hauglin (1996) for details.

(2) Direct juxtaposition
realfag
real+fag
natural science + mathematics

(3) Epenthesis

a. gårdshund
gård+hund
farm dog

b. sauebonde
sau+bonde
sheep farmer

(4) Hyphen
ABC-våpen
ABC+våpen
ABC weapon

(5) Suppletive stem
klesklype
klær+klype
clothes pin

(6) Spelling changes
busstasjon
buss+stasjon
bus station

When three or more identical letters are adjacent, only two are written, as shown in example 6 above. Furthermore, many Norwegian words can be spelled in more than one way, for instance the compound *andregradsligning* (second-degree equation) can also be spelled *andregradslikning*, *annengradsligning* and *annengradslikning*.

Finally, when more than two stems are involved, the compound can technically be split more than one way. Example 1e), *storgårdshus*, can be split as

(7) a. (stor+gård)+hus
house of a large farm

b. * stor+(gård+hus)
large farmhouse

The split in example 7b) is not reasonable because the epenthetic *s* tends to force a split when ambiguity arises.

An example like 1d) splits to ((gård+hus)+(tak+stein))+problem since both *gårdshus* and *takstein* functions as stems of their own.

2.2 Compounding in English

Compounding in English is not so clear cut. It is more fruitful to look at what Norwegian compounds translate to in English[2]:

(8) noun + noun
  a. gårdshus - farmhouse
  b. bygård - apartment building

(9) adjective + noun
  a. statsvitenskap - political science
  b. forretningsstrøk - commercial area

(10) noun + *of*-phrase
  a. produksjonsmidler - means of production
  b. streikevarsel - notice of a/the strike
  c. allemannsrett - public right of access

(11) single words
  a. arbeidsgiver - employer
  b. hovedstad - capitol

(12) other
  a. arbeidsro - opportunity to work undisturbed
  b. skoleplikt - compulsory school attendance
  c. skolevei - way to and from school
  d. streikerett - right to strike
  e. kjøpelyst - desire to buy

Of the examples 8 to 12, only example 8a) wasn't found in any dictionary, while the examples 10c), 12a) and 12e) were found in the bilingual dictionary but not in Bokmålsordboka.

3 The prototype

The entire recompounding process is shown in figure 1.

At point a), after having been handed an already lemmatized word, `ReCompounder` first tries to look up the word in its bilingual dictionary. If the word is not found it tries to treat the source-language term as a compound by attempting to split the potential compound into stems. If this is possible, each stem is translated into the target language at point b). The translated stems are then recombined first with each other in point c), then with the templates of point d) into possible compounds of the target language at point e). Finally, `ReCompounder` checks whether these potential compounds are in use on the web at point f), and from the result of that test, in point g), the most frequent candidate is chosen as the best translation.

The assumptions are as follows:

- The stems in a compound each carry their most frequent meanings
- The translations/meanings in the bilingual dictionary are sorted by frequency, most frequent first
- It is sufficient to split a compound only once, into just two stems, because if a compound stem is used in a larger compound, it carries only its most frequent meaning
- If a compound exists in Norwegian there is a construction in English with approximately the same meaning that already exists on the indexed part of the web

3.1 Use of the dictionaries

The prototype searches through a digitized version of the Engelsk Stor Ordbok, treating it as a treebank. Furthermore, the number of known Norwegian stems have been increased by the stems in the NorKompLeks-lists, and checked against Bokmålsordboka.

3.2 The compound splitter/recognizer

The compound splitter[3] at point a), figure 1 works by attempting to split the word in two. Each piece, hereafter called the *original stems*, is then looked up in a list of known stems, so as to filter out misspelled stems.

3.3 The stem-translator

After having been split, each each original stem is looked up in the bilingual dictionary at point b). The translations of each are then stored in a single ordered, duplicate-free list per original stem, as in example 13.

(13) gård: $\{farm, estate, \ldots\}$
  hus: $\{house, building, \ldots\}$

These lists of *stem candidates* are then stored in the order they were made into another ordered list, thereby maintaining the order of the original stems, as in example 14.

[2] Most of the examples were found in Hasselgård et. al. (1998).

[3] The compound splitter can also be used as a compound recognizer

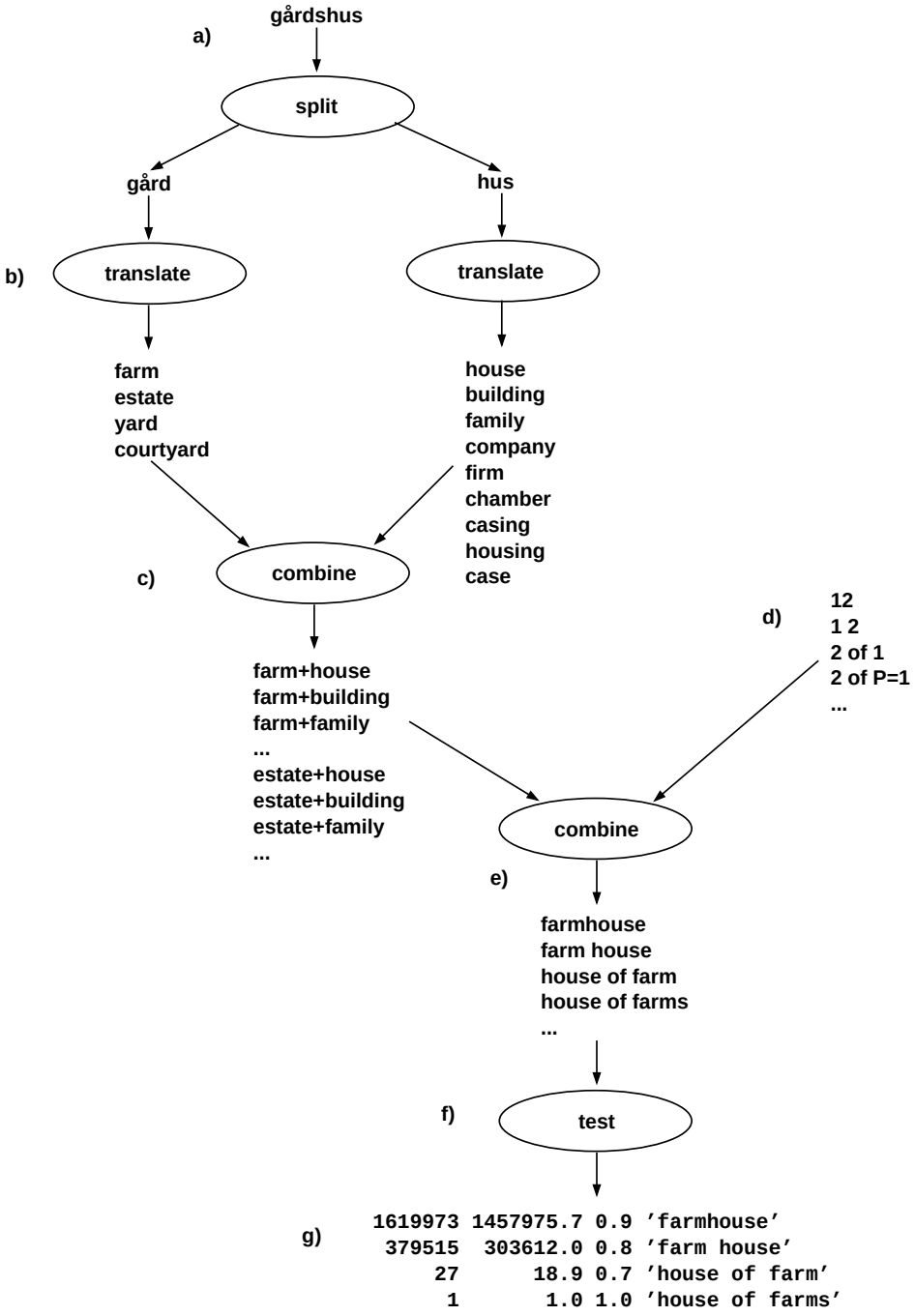

Figure 1: The recompounding process from a), the splitting of the compound, to g), the result from the web search.

(14) $\{\{farm, estate, \ldots\}, \{house, building, \ldots\}\}$

3.4 Potential translation candidates

The *translation candidates* are the strings to be tested with an Internet search engine. They are derived by first combining each stem candidate for each position in the original compound while maintaining the order of the original stems, as shown in point c) of figure 1, then combining these with the *templates* from point d), resulting in the translation candidates at point e).

In table 1, the examples 8 to 10 have been used to show how the templates bridge the translation between the Norwegian and English words.

Norwegian	Template	English
gård$_1$+hus$_2$	12	farm$_1$house$_2$
by$_1$+gård$_2$	1 2	apartment$_1$ building$_2$
forretning$_1$+strøk$_2$	A=1 2	commercial$_1$ area$_2$
produksjon$_1$+middel$_2$	P=2 of 1	means$_2$ of production$_1$
streik$_1$+rett$_2$	2 to 1	right$_2$ to strike$_1$

Table 1: Each digit in the template is replaced by the correspondingly indexed stem candidate. A= means the following stem is an adjective derived from the indexed noun, P= means the following noun-stem must be in the plural.

3.5 Testing the candidates

Each translation candidate is then tested by doing a web search on Google or Yahoo!, at point f), yielding a list like the one at point g). The first column is the raw frequency as extracted from the search, the second is the frequency adjusted for quality, the third is the quality and the fourth is the translation candidate.

Adjusting for quality is needed due to the limits mentioned in section 3.7. The first ten results of each run are checked to see if they actually contain the search terms in the required order and without intruding punctuation. The number of results that *do* contain the search terms in the returned context is then divided by ten, yielding the *quality* of the search, a number between 0.0 and 1.0, inclusive. Assuming that following pages of results have the same ratio of good hits to bad[4]. The total frequency is then multiplied with the quality, yielding the *adjusted frequency*. To date, this has not changed the order of the two highest scoring candidates.

3.6 Other oddities

The prototype has been used to experiment with *snowclones* (Pullum, 2004), a variety of cliched fixed expressions where one or more specific words can be replaced by words that look almost the same written, sound almost the same spoken, or belong to the same part of speech. One example is "All your base are belong to us.", a catchphrase that still has over 500000 hits on Google. Variations include "All your base are belong to U.S", "All your bias are belong to us" and "All your office are belong to me".

If the target language has a construction with similar meaning and same number and types of replaceable parts, these can be translated in the same vein as `ReCompounder` translates compounds. However, since there are few if any snowclones in the texts in the domain of the LOGON project, further experiments have been put on hold.

3.7 Search engine capabilities

Due to consolidation and mergers in the search engine industry, as of today the two leading search engines, both in terms of users and index-size, are Google at `http://www.google.com/` and Yahoo!'s search engine at `http://search.yahoo.com/`. *Alta Vista*, which was previously used for a similar purpose (Grefenstette, 1999), is now a front-end to Yahoo!, as is *Alltheweb* (personal communication with Yahoo!). Not only are the indexes much larger[5], but the query syntax and ranking algorithms are also different from 1999. In addition, instead of parsing a web page of results the search engines can now be accessed and used through stable APIs (Application Program Interface).

3.7.1 Number of queries

When this paper was written, the Google-API had a limit of 1000 queries per developer id per undefined day. The limits for Yahoo! were 5000 queries per application id per ip address per 24 hour period. An exhaustive search for a translation of *gårdshus* costs 200 queries,

[4]The number of bad hits generally increases, but does so depending both on filtering done by the particular search engine as well as the number of hits.

[5]As of September 27, 2005, index-size is no longer given on the search-pages (Sullivan, 2005).

meaning there's a very limited amount of exhaustive searches possible per "day". Using a page scraper[6] would get around the limits but the search engine providers provided the APIs so that various web-robots do not count as people for statistical purposes.

3.7.2 Possible queries

Currently, it is not possible to use web search to discover whether an English compound is written with or without a hyphen due to how "hyphen" is interpreted by the search-engines. When this paper was written, a search for `"stem1-stem2"` now yields approximately the same result as a search for `"stem1 stem2"` `stem1stem2` (Liberman, 2005) if using Google.

3.7.3 Possible results

Punctuation is completely ignored, so a search for rare phrases like `"repair truck"` might return only results containing the two words separated by punctuation. The hits might not even contain the exact search terms at all, so it is always necessary to check if the returned pages actually contains the query, especially if the number of returned pages is in the low hundreds.

3.7.4 Precision and recall

Web search excels at recall. If there is a document in the index that contains one of the words of the search query, or a link to a document that is not indexed but was found to contain one of the words of the search query, or it fulfills other criteria only known to the designers of the search engine in question, a link to the document will be returned. As can be seen from the previous paragraph, these criteria doesn't necessarily lead to the same results as what a language researcher would call a relevant hit.

Furthermore, the algorithms for how the hits are ranked are not known and are subject to change, as does the actual number of total documents indexed, the number of documents containing information relevant to the query, how the query is interpreted, how the relevancy is computed, how the frequency given is estimated from the actual frequency and how the data is presented to the consumer. Ergo, the traditional formulas for precision and recall cannot be used.

[6] Page scraping: parsing a page directly, synonyms: web scraping, site scraping, screen scraping

The web search presents you to a snapshot of the document index and query syntax and ranking algorithm at the moment of search. A later search might not return the same results, the same ranking or the same frequency, but the relative frequency of the query as compared to another query changes only when the index is updated, thus slowly. This means that the differences between the frequencies of the results of many searches done in a short and limited period of time will be approximately the same, and that is the measure used here.

3.7.5 Linguistic limits

There is no guarantee that the most frequent candidate translation is the *best* translation. The system as it stands does no domain checking what so ever, using the entire index for its searches. The readers and writers of documents on the web are the subset of people that can afford to and have the time to access it, there is thus a systematic bias. A document might have been written by someone with only a basic grasp of the language.

4 Evaluation

	Count	%of nouns
Nouns, total	228	100.0%
Known	151	66.2%
Unknown, not recompoundable	20	8.8%
Unknown, recompoundable	57	25.0%

Table 2: Overview of the nouns in the corpus.

The prototype was tested on a small corpus built by Ola Huseth in January 2004. It is a collection of tourist information built from web pages that had both a Norwegian and an English version. Stylistically they resemble advertisements, having a high frequency of adjectives per sentence.

The evaluation corpus consisted of 112 sentences including fragments. A locally made postagger was used to extract the common nouns, finding 263 potential, inflected, nouns. While the tagger is capable of returning uninflected forms of known words, compounds are generally unknown, so a basic lemmatizer based on the same principle as the compound splitter was made to derive the base forms of the nouns.

The lemmatizer rejected 21 of the tagged nouns as they did not end with an in-

flected noun or noun stem[7], another 14 words mistagged as nouns was found by manual inspection. Of the remaining, 151 already had translations in the dictionary and was discarded. At this point, duplicates were removed, leaving 77 potential compounds to be translated by ReCompounder.

ReCompounder had translation suggestions for 57 of the remaining words.

While most of the failures was due to misspellings and mistaggings that had not been caught earlier or the original stems not being in the bilingual dictionary, there were two very interesting instances of a related, non-software problem. The word *bautastein* could not be translated because the original stem *bauta* translates to "menhir", a menhir is a standing stone, and there were no instances of `menhir stone` or with other synonyms for stone on the web. This is a good indication that the word `bautastein` itself needs to be added to the bilingual dictionary. A similar case was *fiskevær*, meaning "fishing village". One of the meanings of the stem *vær* is also "fishing village", ergo one wound up with a reduplicated stem in the candidate.

	Random	Ranked
Highest ranked is good	7.0%	36.9%
At least one good	21.1%	31.6%
At least one possible	33.3%	17.5%
No good candidates	38.6%	14.0%

Table 3: Evaluation-results: random sample of size 5 from all possible combinations and five best existing collocations sorted by score

Evaluation was done by having a human bilingual in Norwegian and English rate the suggestions. For each list of translation candidates, the evaluator was presented with a context for the original compound and with up to five of the best candidates[8], as seen in figure 2. The evaluator was then asked to quickly mark each candidate as to whether it was an acceptable translation, completely unacceptable or potentially acceptable. This is summarized in the column *Ranked* in table 3.

In addition to the sets of ranked candidates the evaluator was also given sets of candidates that were random samples of all generated, unchecked candidates, to serve as a basis for comparison. This is the column *Random* in table 3.

The most desirable outcome is that the candidate of the highest frequency is suitable as a translation, as shown in the second row of table 3. More interestingly, there was at least one good translation out of five about two thirds of the time, as shown by the second and third rows.

5 Results so far

Already while testing the system during the development stage, certain tendencies emerged. All the compounds so far tested can be categorized as follows:

Good to excellent translations

(15) gårdshus → gård + hus
$\Longrightarrow$ farm house → farmhouse

The top five of the complete results for farmhouse are in table 4.

1427890	1285101.0	0.9	farmhouse
332322	265857.6	0.8	farm house
72417	50691.9	0.7	farm family
35695	24986.5	0.7	estate family
34652	31186.8	0.9	farm building

Table 4: Top five results for “farmhouse”. Notice how “farm building” beats “estate family” when ranking by adjusted frequency.

Unsplittable compound No stems to search, no result. Often this is due to a compound-only suppletive stem missing from the dictionary or wordlists, as discussed in section 2.1, example 5.

Bilingual dictionary lacks usable English stem Sometimes, this is due to the stem missing from the dictionary, but there are also cases where there are no single word translations of the stem.

(16) seterbu → seter + bu
$\Longrightarrow$ no useful translations + ...

In the above case, there are no suitable one-word translations of *seter*. This is different from the last category in that here the target language lacks an adequate term, while the last category is when the source language term lacks an entry for the adequate meaning.

[7] and thus not translatable in the current system

[8] Sorted by score, but the evaluator was not made aware of this.

```
46. 604 meter over Lysefjorden henger denne markante fjellformasjonen.

fjellformasjon

[ ] mountain formation
[ ] formation of mountains
[ ] formation of mountain
[ ] formation of mount
[ ] mount formation
```

Figure 2: Evaluation

One of the stems of the compound means the same as the compound This was the case with *bautastein* and *fiskevær* as discussed in section 4.

The meaning of the Norwegian compound stem systematically differs from stand-alone stem

(17) ulvestamme → ulv + stamme
$\Longrightarrow$ wolf + no useful translations

This is so far the most interesting result, showing that the meaning of the stem when part of a compound is missing from the dictionary. <mammal> + *stamme* means approximately "total population of <mammal> in a certain area" and is this synonymous with *bestand*. A similar problem is *bil*, when in a compound it generally translates to “truck”, not “car”. While ReCompounder can be used some of the way to discover such holes in the dictionary, making the definition of the missing stem or meaning still takes a lexicographer.

6 Conclusion

Using current web search engines instead of traditional static corpora is a workable strategy for automatic translation of compounds.

Compared to a baseline where all combinations are equally likely, the results are pretty good. As a bonus, the system as-is detects errors and gaps in the bilingual dictionary and is therefore useful as a tool for lexicography.

The strategy itself will be used in the machine-translation project LOGON to improve coverage there.

References

Øystein Eek et al., editors. 2001. *Engelsk stor ordbok: engelsk-norsk/norsk-engelsk*. Kunnskapsforlaget, Norway.

Gregory Grefenstette. 1999. The WWW as a resource for example-based machine translation tasks. In *Proceedings of the ASLIB Conference on Translating and the Computer*, London, October.

Hilde Hasselgård, Stig Johansson, and Per Lysvåg. 1998. *English Grammar: Theory and Use*. Universitetsforlaget, Oslo, Norway.

Janne Bondi Johannesen and Helge Hauglin. 1996. An Automatic analysis of compounds. In T. Haukioja, editor, *Papers from the 16th Scandinavian Conference of Linguistics*, pages 209–220, Turku/Åbo, Finland.

Mark Liberman. 2005. Google usage. Language Log, `http://itre.cis.upenn.edu/~myl/languagelog/archives/001956.html`, 6 March.

Jan Tore Lønning, Stephan Oepen, Dorothee Beermann, Lars Hellan, John Carroll, Helge Dyvik, Dan Flickinger, Janne Bondi Johannsen, Paul Meurer, Torbjørn Nordgård, Victoria Rosén, and Erik Velldal. 2004. LOGON. A Norwegian MT effort. In *Proceedings of the Workshop in Recent Advances in Scandinavian Machine Translation*, Uppsala, Sweden.

Torbjørn Nordgård. 1996. NorKompLeks: Some Linguistic Specifications and Applications. In Lindebjerg, Ore, and Reigem, editors, *ALLC-ACH ’96. Abstracts*, Bergen. Universitetet i Bergen, Humanistisk Datasenter.

Geoffrey K. Pullum. 2004. Snowclones: lexicographical dating to the second. Language Log, `http://itre.cis.upenn.edu/~myl/languagelog/archives/000350.html`, 6 January.

Ulrike Rackow, Ido Dagan, and Ulrike Schwall. 1992. Automatic translation of noun compounds. In *Proceedings of COLING 92*, pages 1249–1253, Nantes, France, August.

Danny Sullivan. 2005. End Of Size Wars? Google Says Most Comprehensive But

Drops Home Page Count. Search Engine Watch, http://searchenginewatch.com/searchday/article.php/3551586, 27 September.

Takaaki Tanaka and Timothy Baldwin. 2003. Translation selection for Japanese-English Noun-Noun Compounds. In *Proc. of Machine Translation Summit IX*, pages 378–385, New Orleans, USA.

Boye Wangensteen, editor. 2005. *Bokmålsordboka*. Kunnskapsforlaget, Norway.