Corpora in machine translation

Hanne Moa

February 2, 2005

1 Introduction

> "Text must be (minimally) understood before translation can proceed effectively. Computer understanding of text is too difficult. Therefore, Machine Translation is infeasible." (Bar-Hillel, 1960)

In spite of this quote there are many machine translation systems in use today, and more are being made, as the need for translations is seemingly boundless. For instance the contracts, agreements, laws and parliamentary sessions of the EU need to be translated somehow, and even if machine translation as good as human translation is infeasible, as that is what Bar-Hillel was concerned about, even quickly made, partial and rudimentary translations can be of help to a translator, or to choose what texts need to be properly translated in the first place.

Intriguingly, it turns out that corpora can help with the second claim of the quote, that "Computer understanding of text is too difficult.". Corpora can provide some understanding of the world simply by being a source for deriving frequencies or other statistically significant phenomena like finding collocations and words that don't follow the rules.

In this paper I will zoom in from the general to the specific, going from the past up to today. Section 2 looks at early use of corpora and machine translation (hereafter MT), section 3 is about modern MT and its growing dependence on corpus linguistics and finally, section 4 is about a specific MT-project that is still under development and its use of corpora, namely LOGON (Lønning et al., 2004).

2 A brief history of MT and corpus research

This little summary of history concentrates on the years before the 1990s, as that decade saw the explosion in popularity of corpus linguistics and especially corpus-based machine translation in the guise of Statistical-Based MT (SBMT), which is described on page 8.

For MT, a short but detailed historical overview can be found in Jurafsky and Martin (2000, chapter 21), for even more depth see Slocum (1985). As for corpus linguistics, the main source has been chapter 1 of McEnery and Wilson (2001), an overview of the earliest period can be found in in chapters 1.2-1.3 of that book while the later years are detailed in chapter 1.5.

2.1 Corpora before computers

Since one cannot really talk about Machine Translation without the machines, describing corpora before the machines is equally unhelpful. Suffice to say, prior to the paradigm-change initiated by Noam Chomsky in the late 1950s, most if not all linguistic research started out with amassing a corpus of the language of interest, then building theories from the observed data or checking if theories fit with the observed data.

According to Chomsky, the problem of using corpora is that language is infinite in its complexity while corpora will always be finite, hence everything explained by a corpus will be biased by the contents of that corpus.

After Chomsky, the use of corpora went underground except in fields like phonetics and language acquisition where introspection was less useful: the sounds of language can be rigorously collected and studied and a child who still cannot string together a whole sentence cannot be expected to be able to understand the difference between nouns and verbs or whether something sounds odd.

Luckily, while the linguists started using mainly themselves as informants, computers and computing had been out-growing the government bunkers for some time and were looking for new and interesting things to compute, eventually giving rise to corpus linguistics as it is today, and this is further described in the next section.

2.2 Corpora during the cold war

While corpora were out of fashion in linguistics it lived on elsewhere, and as computer technology improved, were increasinly stored, searched and analyzed on computers.

Corpora in the humanities Resarchers of ancient texts have always used a more or less corpus-based approach, and they were quick to start using computers for their concordances when such were available. In fact, as far as is known, the very first machine-readable and -searchable corpus was made for Father Roberto Busa's card-based collection of texts by St. Thomas of Aquinas.

Mechanolinguistics Mechanolinguistics was the proto-form of corpus linguistics as we know it today, and it started in France in the 1950s. The French linguist Alphonse Juilland encountered and developed solutions to the problem of finite data describing infinite language, like which texts to select for a corpus to decrease bias and how to annotate.

Corpora to study of the English language The use of coprpora for studying English did not die out, as shown by work on the Brown-corpus starting in 1960. During the 1970s, corpora for English were computerized and it is this branch of corpus linguistics that would eventually produce corpora like the Lancaster-Oslor-Bergen corpus (LOB) and the British National Corpus (BNC).

The legacy of Firth One of the giants of British linguistics was John R. Firth, and he preferred studying what ordinary people actually said and wrote, using corpora. Firth is also the source of terms like *collocation* (see section 3.1).

The COBUILD project and its corpus, the Bank of English, was built by people schooled in the Firthian tradition, which differs from other corpus linguistics in that it uses open-ended corpora containing whole texts instead of fragments selected to lower bias.

2.3 MT sans corpora

The idea of automatic translation by computers surfaced already at the end of the 1940s (Weaver, 1955) and at the end of the 1950s several projects were underway. However, in 1960 a highly critical article by Bar-Hillel was published (Bar-Hillel, 1960), among other things containing the quote in the introduction, and shortly thereafter, as with corpora, MT went underground.

The first experience many have with machine translation today is through Altavista's Babelfish[1], which uses *systran*. *systran* is an old system: it was first brought into use in 1970 (Slocum, 1985), replacing an older system at the USAF for translating from Russian to English, and has been refined ever since. *systran* is a direct[2] translating Rule-Based[3] system.

In 1976 the weather report-translating system Météo was implemented, it used "Q-systems", an early form of unification, and was thus laying the foundations for constraint-based systems.

There was a rebound in the late seventies and the MT of this time was heavily influnced by concepts from Artificial Intelligence[4]. Where systems like *systran* are described by the lower third of the MT pyramid in figure 1, these later systems belong to the upper part of the pyramid, depending on more or less in-depth analysis of the syntax and semantics of both languages to be translated in order to make transfer-rules, or to design an interlingua.

Finally, in the early 1990s generally available computers were sufficiently fast and capable of storing sufficiently large amounts of data that both corpus linguistics as we know it today, and corpus-based methods of machine translation were made possible, and that is the topic of the next section.

3 Current trends in Machine Translation

"You shall know a word by the company it keeps." (Firth, 1957)

The lack of world-experience is still the main problem for MT, and a way to overcome this is to specialize the translation-system for a particular, narrow and well-defined domain like weather reports, which needs little world-knowlegde. Another is to add world-knowledge to the system somehow, whether explicit as with Knowledge-Based MT (KBMT), implicit through using corpora as with Statistical-Based MT (SBMT) or by trusting a human to impart it where needed, as in Dialogue-Based MT (DBMT).

Before we embark on describing the various paradigms I will define some of the MT-specific terms that are used throughout the paper.

[1] See `http://world.altavista.com/` (accessed 2005-01-08)

[2] See figure 1

[3] See section 3.2.1 on page 8.

[4] For an AI view on natural language (as opposed to artificial languages and programming languages) see Luger and Stubblefield (1998, chaper 11).

3.1 Some relevant terms

The MT pyramid (figure 1) is one way of visualizing different ways of translation. It goes from a minimal amount of analysis at the bottom through more or less deep analysis for transfer in the middle to the creation of an intermediary language at the top.

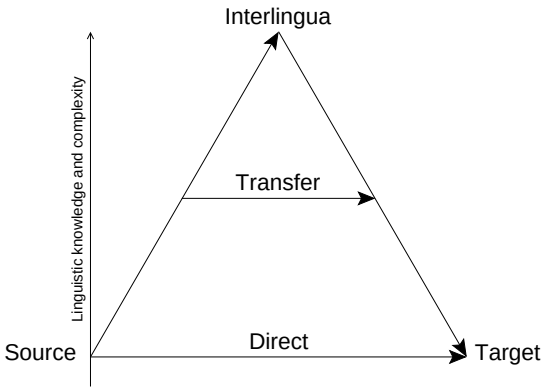

Figure 1: The MT pyramid

Direct translation tries to get away with as little analysis as possible. Word for word translations are direct, and if two languages are similar enough, this might be sufficient.

Transfer covers the ground between direct translation and translation via interlingua, and utilizes analyses of at least the syntax but also often the semantics of the languages in question. An often heard objection to transfer is that you need a set of transfer-rules for every language pair. See figure 2.

Interlingua tries to overcome the need for rules for each language pair, instead each language is transfered to and from the interlingua. However, this often means the interlingua needs to be a super-set of the languages to be translated, and furthermore going via an interlingua tends to bleach away the style and tone of the original.

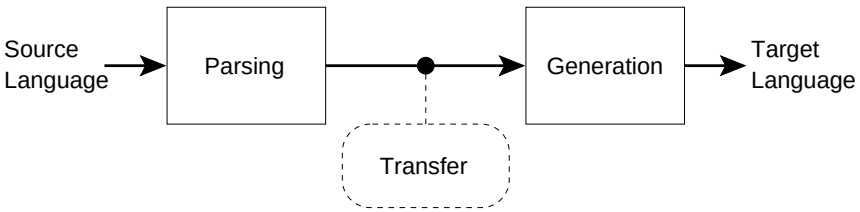

Figure 2: Transfer

Collocations aka. multi-word units (MWU) As a search through the literature will show, there are still no agreed upon definitions for collocations and multi-word units (hereafter MWUs), or which one of them includes sayings, titles, idioms or phrasal verbs or which one contains the other. Prior to the use of corpora went underground in the 1950s however, collocations were of high importance and were defined by John Firth (1957) thusly: "Collocations of a given word are statements of the habitual or customary places of that word." For an overview of techniques for discovering collocations in corpora, see Manning and Schütze (1999, chaper 5).

In this text, I use MWU to mean both collocations and MWUs, and define an MWU to be *any group of two or more words that cannot be handled by word-based syntactical or semantical rules alone but might have to be looked up or otherwise need their own specific rules.* This handily includes only the subset of MWUs possible, specifically only the ones that need to be considered in a system of machine translation. There will after all be cases where a (wider definition) MWU in one language can be translated by conventional means to another language, e.g. when one of the languages has borrowed an idiom or saying from the other.

A last problem for the definition of MWUs is whether compounds are MWUs or not, and what a compound *is*. In e.g. English most new compounds are written with spaces between each stem, and it is debated whether "blackbird" in one word is a compound or not, and whether an actual "black bird", adjective+noun, is a compound. In other languages like German or Norwegian, compounds are to be written without spaces and "en svart fugl" ("a black bird") is never a compound. Going the other way, in Norwegian the compound "gråspurv" ("English sparrow") contrasts with "grå spurv", "grey sparrow", which is not considered to be a compound. A possible solution to this is to define compound as it is valid per language pair in the system, and then see whether they need to be special-cased or not, thereby deciding whether or not they should be considered to be MWUs.

Alignment and extraction Alignment is matching up pieces of one thing with equivalent pieces of another, in the case of natural language processing this usually means matching pieces of text with other pieces of text. Most relevant for MT is the alignment of sentences, phrases and words or a spoken word to a written word.

Alignment is maybe the most important part of SBMT (see page 8) as the training texts needs to be well aligned before translation can take place.

As for definitions, alignment...

> "Alignment is the process by which correspondences of sub-components within the paired sentences are found. The goal is to find as many correspondences as possible for *some* substructure in the parallel corpus." (Yamamoto and Matsumoto, 2003)

should not be confused with the very similar *extraction*...

> "Extraction, as the name suggests, focuses on extracting subcomponents that correspond by processing the entire parallel corpus. The goal is to find correspondences for *some* substructure in the parallel corpus." (Yamamoto and Matsumoto, 2003)

which is very important for EBMT.

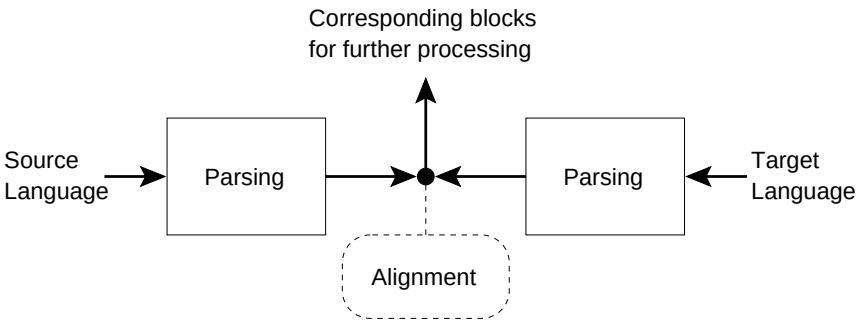

Figure 3: An abstracted overview of alignment, compare it to figure 2 of transfer.

Minimal Recursion Semantics (MRS) While MRS (Copestake et al., 1995) is a way of encoding semantics, lower-case "mrs" is used in the LOGON-project to designate an actual semantic encoding of a clause. The sentence in 1a thus has the mrs of 1b, and as example 1b shows, in LOGON they are stored as attribute-value matrixes.

(1) a. Været var flott.
weather.the was great
'the weather was great'

b.
$$\begin{bmatrix} \text{TOP} & \text{h1} \\ \text{INDEX} & \text{e8} \\ \text{RELS} & \left\{ \begin{bmatrix} \text{stative\_asp\_rel} & \\ \text{LBL} & \text{h7} \\ \text{ARG1} & \text{e8} \end{bmatrix}, \begin{bmatrix} \text{prpstn\_m\_rel} & \\ \text{LBL} & \text{h1} \\ \text{MARG} & \text{h9} \end{bmatrix}, \begin{bmatrix} \text{\_flott\_j\_1\_rel} & \\ \text{LBL} & \text{h7} \\ \text{ARG0} & \text{e8} \\ \text{ARG1} & \text{x5} \\ \text{LNK} & 20 \end{bmatrix}, \right. \\ & \left. \begin{bmatrix} \text{def\_rel} & \\ \text{LBL} & \text{h3} \\ \text{ARG0} & \text{x5} \\ \text{BODY} & \text{h4} \\ \text{RSTR} & \text{h2} \\ \text{LNK} & 2 \end{bmatrix}, \begin{bmatrix} \text{\_vær-met\_n\_rel} & \\ \text{LBL} & \text{h6} \\ \text{ARG0} & \text{x5} \\ \text{LNK} & 2 \end{bmatrix} \right\} \\ \text{HCONS} & \left\{ \text{h9 QEQ h7, h2 QEQ h6} \right\} \end{bmatrix}$$

3.2 An overview of current paradigms in MT and their use of corpora

Large corpora is a relatively new phenomenon, as it depends on fast computers and large amounts of storage-space. Many of the MT-paradigms below started out before such computers were generally available and therefore have had to make do without large corpora, but it can be safely assumed that a bilingual

dictionary, be it on paper or on a hard disk, has played a role in their construction. Furthermore, several of the paradigms lead to the construction of corpora that might be useful for other uses, e.g. LBMT builds up a bilingual lexicon that also includes a large amount of just the MWUs that need special attention when translating.

This section is mostly a summary of Dorr et al. (1998, section 5) except for my musings on corpora-use and wherever there are explicit citations.

3.2.1 Linguistics based MT

The older paradigms of MT, like rule-based MT, are all based on more or less linguistic analysis of the languages to be translated. Less analysis means direct translation, more leads to translation by transfer or via an interlingua.

Constraint-Based MT (CBMT) is generally transfer between unification-based grammars since it is relatively easy to express constraints through unification. An attribute-value matrix (AVM) of the source-sentence is transfered to a valid, equivalent AVM in the target-language, and the translation of the source-sentence is generated from this AVM.

Recent examples include Verbmobil (Emele et al., 2000), which aimed to translate speech between English, German and Japanese through deep semantic transfer. A project still in progress is LOGON Lønning et al. (2004) which will be covered in section 4.

Knowledge-Based MT (KBMT) is related to Knowledge Representation (KR) (Luger and Stubblefield, 1998, chapter 8) from AI. KBMT models the world through ontologies and then uses the ontologies to provide an understanding of the text when translating. As the weight so far has been on building the ontologies and not the translation itself such systems are awkwardly linked with syntactic phenomena. Another problem inherent to all ontologies is that they are models of cooccurences and connections in the world, and different people, cultures and languages draw the connections differently. This leads to the same problem that plagues systems based on interlinguas; the ontology needs to be a superset of the world-views involved and in translation, the flair and style of the source text will not be present in the translated text.

Nevertheless, provided that the domain is narrow and well-defined, KBMT will produce fully automated, high-quality translations. For some examples see CyC[5] (Lenat and Guha, 1990; Lenat, 1995) or "topic maps" (Pepper, 2002).

Lexical-Based MT (LBMT[6]) relates the lexical-entries of different languages via rules. The lexical entries in question are per meaning, not per word, so that what is expressed with a single word in one language and an MWU in another is correctly looked up, see for instance example 2 below:

(2) *aimer* $\Leftrightarrow$ *be fond of*

[5]The name is derived from en*cyc*lopaedia. Homepage at `http://www.cyc.com/` (accessed 2005-01-08).

[6]LBMT is also an abbreviation for linguistic-based MT in general.

LBMT becomes in essence a bilingual dictionary with many entries for MWUs. It overlaps with RBMT, PBMT and S&BMT and the techniques involved can be used for the lexicon-handling part of an MT-system.

Rule-Based MT (RBMT) is translation by rules, and often through direct translation as opposed to by transfer or interlingua. RBMT differs from LBMT in that the needs of individual lexical entries are covered imore directly by the grammar, not the lexicon.

One example of a rule-based system is the *Rosetta* system. In *Rosetta*, the head-switching phenomena like in example 3, the English example with the meaning "by chance" as a verb is taken to be the canonical form, and is linked up with the Dutch adverb "toewallig", which is marked as a deviant form. When translating, the deviant mark triggers a run through a head-switching module for the problematic pair.

(3) English: Mary happened to come
Dutch: Mary kwam toevallig
'Mary came by chance'

Principle-Based MT (PBMT) is an alternative to RBMT, where instead of potentially very specialized rules one utilizes a small set of more general principles. Even phenomena like passivization is not given its own rule but is instead the result of several underlying principles.

Like KBMT and EBMT, PBMT can provide good coverage of phenomena, but demands a narrow domain.

Shake and Bake MT (S&BMT) The origin of S&BMT can be traced to two papers from 1992, Beaven (1992) and Whitelock (1992), and it is thus one of the newest approaches. It works by trying all possible target-language words (the "shake") in all possible orders until a sentence is found that satisfies all syntactic constraints (the "bake"). A problem with this is that the numbers of sentences to "bake" grows exponentially depending on the number of words in the sentence times the number meanings of these words, meaning that S&BMT generation is NP-complete (Brew, 1992).

3.2.2 Non-linguistic MT

How can MT not be linguistic? When little to no linguistic analysis is done of the languages involved and one instead depends heavily upon corpora of already existing translations.

Statistical-Based MT (SBMT or SMT) In 1990, something new entered the arena of MT. This was the year that Brown et al. first published about their MT project that *only* utilized bilingual corpora and a surface analysis thereof.

These so-called IBM-models of translation need enormous amounts of bi-texts to train themselves, and the first corpus used for the purpose was the official records of the Canadian parliament, the *Hansards*, being bilingual in English and French.

Since SBMT is wholly dependent on corpora, these need not only be better than just good, but more important they need to be exhaustive, so it is not an option for languages with little if no digitized text. Another problem is that a good alignment, as defined in section 3.1 on page 5, is essential for good results, and this is not a trivial problem to solve, especially as one also need to consider how to align MWUs and whether the alignment should go both ways. A third problem is that to actually translate it is necessary to estimate[7] several language-dependent paramaters, which is the topic of Brown et al. (1993) which is considered the seminal paper on the paradigm. A fourth problem is that SBMT cannot handle long distance dependencies, which is the forte of CBMT and a focus of systems based on HPSG (Pollard and Sag, 1994; Sag et al., 2003). Yet another problem is what to do with languages with heavier emphasis on morphosyntax like Turkish, since the morphological analysis necessary is not done.

SBMT systems need and/or produce sentence- and word-aligned corpora as part of the training process, and these can later be used as concordances for making for instance dictionaries.

For more on SBMT and alignment, I recommend Manning and Schütze (1999, chapter 13) for the general overview and Knight (1999) for an explanation and walkthrough of Brown et al. (1993).

Example-Based MT (EBMT) is *translation by analogy*. A previously unseen sentence or phrase is compared with an already translated sentence or phrase. The approach resembles Case-Based Reasoning from Artificial Intelligence (Luger and Stubblefield, 1998, section 6.4) and is therefore also known as Case-Based MT.

EBMT needs a preexisting database of at least pos-tagged parallel translations and a thesaurus or ontology to compare the similarity of the content-words. The corpora can be much smaller than those needed in SBMT as what is needed is good coverage of syntactic and semantic differences and not vocabulary, though the amount of data is still so large that efficient search is still an unsolved problem[8]. So far, EBMT has mostly been used to translate phrases like for instance noun-phrases between Japanese and English, where there are a few clearly defined patterns. Sentence-translation is possible but since the entire structure of whole sentences then must be stored, it quickly runs into problems with size.

There's a short but fair summary in Dorr et al. (1998)[5.2.2], some current trends can be found in Carl and Way (2003).

Dialogue-Based MT (DBMT) is a lot less automated than the others, aiming to instead be a tool for the human translator. As with EBMT, unseen sentences and phrases in the source language are compared with pairs already translated, but the comparisons are then shown to the human translator which can then adjust and select the best translation, which is

[7] Read: fudge...

[8] Dorr et al. (1998) also mentions difficulty of storage but that was in 1998. With the low prices of large hard-disks these days, a remaining problem is that the hard-disks themselves haven't sufficiently increased access-speed to keep up with the size, thus further hurting search-speed.

stored, ready for comparison with future unseen sentences. As more and more correct, good translations are entered, the system can function more and more autonomously.

DBMT is helped by starting out with a preloaded bilingual dictionary and perhaps some preloaded sentences to start with but this is not necessary. When the domain is large it eventually runs into the same problems as EBMT and KBMT: there are too many sentences and phrases to look up. Therefore it has mostly been used for smaller domains.

Neural Network Based MT (NBMT) is also a recent trend. While neural networks have been used as modules to handle parsing, morphosyntax and disambiguation, using it for fully fledged MT is new. Considering that they can only translate fewer words than there are nodes and that the largest nets still only contain nodes in the hundreds, MT through neural nets can be thought of as no more than proof of concept.

Hybrid systems combine two or more approaches, be they linguistics-based or not. It can be useful to first try a high-quality system and if that fails, go back to a statistics based system so that there always will be a result.

Another way is to use several paradigms simultaneously, letting for instance noun-phrase translation between English and Japanese be handled by an EBMT subsystem while other parts are handled by other, non-EBMT subsystems that is especially suited for their part of the problem.

3.2.3 Indispensible corpora

For all MT systems, corpora are a great aid during evaluation. One selects a representative set of texts from the domain to test with, making sure that this same set was *not* used to train with.

Further discussion on evaluation techniques with the help is beyond the scope of this paper, for an interesting approach through evaluating by comparing machine translations with human translations that is being considered for LOGON, see Papineni et al. (2002).

4 LOGON: a modern MT project

LOGON (Lønning et al., 2004; Oepen et al., 2004) is an MT project that aims to translate text in the tourism-domain from Norwegian to English. More specifically, translating adverts for hiking-trips.

In the core system, the Norwegian is handled by the LFG-based (Dalrymple, 2001; Bresnan, 2001) resource-grammar *NorGram*[9] that is part of the *ParGram*[10] (Butt et al., 1999) project via the XLE[11] software from Xerox. The English is handled by the HPSG-based (Pollard and Sag, 1994) LinGO English Resource Grammar[12] via the LKB[13] (Copestake, 2002) software system.

This system works by transferring mrses[14], which aids in achieving one of

[9]Homepage at `http://www.ling.uib.no/\~victoria/NorGram/` (accsessed 2005-01-09).
[10]Homepage at `http://www2.parc.com/istl/groups/nltt/pargram/` (accessed 2005-01-09).
[11]Homepage at `http://www2.parc.com/istl/groups/nltt/xle/` (accessed 2001-01-09).
[12]Homepage at `http://lingo.stanford.edu/` (accessed 2005-01-09).
[13]Homepage at `http://www.delph-in.net/lkb/` (accessed 2005-01-09).
[14]See section 3.1 on page 6.

the goals of the project: translation that is as similar to the source language as possible, in word order, style and form. In addition there'll be fallback methods to ensure that a translation will always result, even if it is not optimal.

4.1 Corpora used in LOGON

Another of the purposes of LOGON is to create new and improve existing corpora.

4.1.1 The TUR-corpus

One of the meanings of "tur" is "a trip, a hike", and this corpus is a collection of bitexts on hiking by foot in Norway, the domain of LOGON. As of 2004-06-24 this corpus had 360 000 tokens from 536 documents, and it is still being expanded.

The text is collected from tourism websites in Norway, and some of the character of the domain can be seen in example 4. The texts are meant to entice tourists to go hiking so adjectives and intensifiers like "marvellous", "fantastic" and "wonderful" are frequent. The example also demonstrates the variable quality of the translations, 4a being fairly good while 4b is bad to the point of being funny.

(4) a. Her får du en fantastisk utsikt !
Adv V Pron Det Adj N
"You will have a marvellous view!"

b. Mørke skyer i skiftende kontrast med blank og grå himmel er
Adj N Prep Adj N Prep Adj Conj Adj N V
aktørene i et fantastisk skuespill hvor selve Atlanterhavet
N Prep Det Adj N Subj Det N
fungerer som scene .
V Prep N
"Spectacular in winter."

There's a search-interface for the corpus at `http://omilia.uio.no/touristcorpus/` (accessed 2005-01-06).

4.1.2 NorKompLeks

The NorKompLeks project (NKL) (NorKompLeks, 2000) aimed to produce machine-readable computational linguistics oriented dictionaries for the two varieties of written Norwegian, Bokmål and Nynorsk. NKL consists of lists of basic and inflected form of all words in Bokmålsordboka and Nynorskordboka and also contains a rendering of the pronounciation for each word and valency-information for the verbs.

In LOGON these collections see many uses, especially the list of inflected forms as it is often quicker to look up one such than parse it and potentially generate unneccessary ambiguity. Furthermore it acts as a check on and source for vocabulary coverage. An intriguing development has been in using the text-corpora of Norwegian to discover frequencies of inflected forms; that a word *can* be inflected doesn't mean that it ever *is*. Discovering such frequencies cuts down on ambiguity since it lowers the amount of words that needs to be parsed or looked up.

4.1.3 The E-N-E dictionary by Kunnskapsforlaget

The *Engelsk stor ordbok*[15] (Eek et al., 2001) is a bilingual two-way dictionary of English and Norwegian. Since LOGON is to translate from Norwegian to English, only the Norwegian-to-English half has been used so far. This contains 62554, of which 42635 are nouns, 8466 adjectives and 5027 verbs. There are also 472 "words" that only occur in expressions, like

(5) hulter til bulter
"pell-mell, helter-skelter, at sixes and sevens, in a mess"

where neither "hulter" nor "bulter" exists anywhere outside the expression.

4.1.4 Redwoods Treebank

Redwoods[16] (Oepen et al., 2002) is to be a new type of treebank. It aims to avoid the following limitations of existing treebanks, that is: shallow, surface analysis of the texts, inflexible storage formats that can often only be investigated with specialized tools and representations that are static and not kept current with advances in the field.

Redwoods does this by deeper analysis of the sentences through the use of HPSG, data that can be retrieved at varying degrees of granularity and constant updates and adjustments as the field develops.

The initial release consists of 10 000 annotated trees and the necessary tools to create and maintain the treebank.

4.1.5 Other

In addition to the corpora mentioned above there are numerous throw-away frequency lists and pos-tagged texts in use, often derived from already existing corpora.

5 Conclusion

In this paper I have shown a little of the history of corpus linguistics and machine translation and their sometimes common fate.

While we cannot really speak of real corpus linguistics or MT before the advent of computers, embryonic projects were underway as soon as computers started showing up at universities and larger companies. For different reasons, in the 1960s both MT and corpora went out of fashion for a time, but were kept alive by enthusiasts out of view of mainstream linguistics.

Then came the 1990s and with it fast computers capable of storing vast amounts of data, and with that, corpora that contain several millions of words and an MT-paradigm dependent on them, Statistical-Based Machine Translation.

Finally, I have shown that SBMT is not the golden bullet of MT, as the systems being built today are hybrids, but also that corpora still have a role to play in MT.

[15] Publisher's page at `http://www.kunnskapsforlaget.no/kfshop/new/template.jsp?page=product.jsp?ProductId=226` (accessed 2005-01-06).

[16] Homepage at `http://redwoods.stanford.edu/` (accessed 2005-01-06).

References

Bar-Hillel, Yehoshua. 1960. The Present Status of Automatic Translation of Languages. *Advances in Computers* 1:91–163.

Beaven, J. L. 1992. Shake-and-Bake Machine Translation. In *Proceedings of the 14th International Conference on Computational Linguistics*. Nantes, France.

Bresnan, Joan. 2001. *Lexical-functional syntax*. Blackwell Textbooks in Linguistics. Blackwell.

Brew, C. 1992. Letting the Cat out of the Bag: Generation for Shake-and-Bake MT. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, 610–616. Nantes, France.

Brown, P. F., J. Cocke, S. A. Della Pietra, V.J. Della Pietra, F. Jelinek, J. Lafferty, R. L. Mercer, and P. Roossin. 1990. A Statistical Approach to Machine Translation. *Computational Linguistics* 16:79–85.

Brown, P. F., S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics* 19:263–311.

Butt, Miriam, Tracy Holloway King, María-Eugenia Niño, and Frédérique Segond. 1999. A grammar writer's cookbook. Number 95 in CSLI lecture notes. CSLI Publications.

Carl, Michael, and Andy Way, ed. 2003. *Recent Advances in Example-based Machine Translation*. Kluwer.

Copestake, Ann. 2002. *Implementing typed feature structure grammars*. CSLI Lecture Notes. CSLI publications.

Copestake, Ann, Dan Flickinger, Rob Malouf, Susanne Riehemann, and Ivan Sag. 1995. Translation using Minimal Recursion Semantics. In *Proceedings of the 6th. International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-95)*. Leuven, Belgium.

Dalrymple, Mary. 2001. Lexical Functional Grammar. volume 34 of *Syntax and semantics*. Academic Press.

Dorr, Bonnie J., Pamela W. Jordan, and John W. Benoit. 1998. A Survey of Current Paradigms in Machine Translation. In *Advances in Computers*, ed. M. Zelkowitz, volume 49. Academic Press, London.

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

Emele, Martin C., Michael Dorna, Anke Lüdeling, Heike Zinsmeister, and Christian Rohrer. 2000. Semantic-Based Transfer. In *Verbmobil: Foundations of Speech-to-Speech Translation*, ed. Wolfgang Wahlster, Artificial intelligence, 361–379. Springer.

Firth, John R. 1957. A Synopsis of Linguistic Theory 1930-1955. *Studies in Linguistic Analysis* .

Jurafsky, Daniel, and James H. Martin. 2000. *Speech and language processing*. Prentice-Hall, Inc.

Knight, Kevin. 1999. A Statistical MT Tutorial Workbook. URL `http://www.isi.edu/natural-language/mt/wkbk.rtf`, this unpublished document explains Brown et. al 1993, which is rather densely written and heavy on the mathematics.

Lenat, Douglas B. 1995. Cyc: A Large-Scale Investment in Knowledge Infrastructure. *Communications of the ACM* 38.

Lenat, Douglas B., and R. V. Guha. 1990. *Building Large Knowledge-Based Systems*. Reading, MA, USA: Addison-Wesley.

Lønning, Jan Tore, 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*, 6. Uppsala, Sweden. URL `http://stp.ling.uu.se/RASMAT/extended_abstracts/LOGON.pdf`.

Luger, George, and Wiliam Stubblefield. 1998. *Artificial Intelligence. Structures and Strategies for Complex Problem Solving*. Addision Wesley Longman, Inc., 3 edition.

Manning, Christopher D., and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT press.

McEnery, Tony, and Andrew Wilson. 2001. *Corpus linguistics*. Edinburgh: Edinburgh University Press, 2 edition. URL `http://www.ling.lancs.ac.uk/monkey/ihe/linguistics/contents.h%tm`.

NorKompLeks. 2000. NorKompLeks. URL `http://dbh.nsd.uib.no/nfi/rapport/?Keys=7975&language=en`.

Oepen, Stephan, Helge Dyvik, Jan Tore Lønning, Erik Velldal, Dorothee Beermann, John Carroll, Dan Flickinger, Lars Hellan, Janne Bondi Johannessen, Paul Meurer, Torbjørn Nordgård, and Victoria Rosén. 2004. Som å kappete med trollet? Towards MRS-based Norwegian – English Machine Translation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation*. Baltimore, MD. URL `https://mt.uio.no/pub/bscw.cgi/d23044/tmi04.pdf`.

Oepen, Stephan, Kristina Toutanova, Stuart Shieber, Christopher Manning, Dan Flickinger, and Thorsten Brants. 2002. The LinGO Redwoods Treebank: Motivation and Preliminary Applications. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, 1253–1257. Taipei, Taiwan.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for the Computational Linguistics (ACL)*, 311–318. Philadelphia, USA.

Pepper, Steve. 2002. The TAO of Topic Maps: Finding the Way in the Age of Infoglut. URL `http://www.ontopia.net/topicmaps/materials/tao.html`.

Pollard, Carl, and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago, IL, USA: University of Chicago Press.

Sag, Ivan A., Thomas Wasow, and Emily M. Bender. 2003. *Syntactic Theory: A Formal Introduction*. Number 152 in CSLI Lecture Notes. Stanford, CA, USA: CSLI Publications, 2 edition.

Slocum, Jonathan. 1985. A Survey of Machine Translation: Its History, Current Status, and Future Prospects. *Computational Linguistics* 11.

Weaver, W. 1955. Translation. In *Machine Translation of Languages*, ed. W. N. Locke and A. D. Boothe, 15–23. Cambridge, Massachusets, USA: MIT Press. Reprinted from a memorandum written by Weaver in 1949.

Whitelock, P. 1992. Shake-and-Bake Translation. In *Proceedings of the 14th International Conference on Computational Linguistics*, 784–791. Nantes, France.

Yamamoto, Kaoru, and Yuji Matsumoto. 2003. Extracting Translation Knowledge from Parallel Corpora. In *Recent Advances in Example-based Machine Translation*, ed. Michael Carl and Andy Way, chapter 13. Kluwer.