Probing the lexicon in evaluating commercial MT systems

Martin Volk
University of Zurich
Department of Computer Science, Computational Linguistics Group
Winterthurerstr. 190, CH-8057 Zurich
volk@ifi.unizh.ch

Abstract

In the past the evaluation of machine translation systems has focused on single system evaluations because there were only few systems available. But now there are several commercial systems for the same language pair. This requires new methods of comparative evaluation. In the paper we propose a black-box method for comparing the lexical coverage of MT systems. The method is based on lists of words from different frequency classes. It is shown how these word lists can be compiled and used for testing. We also present the results of using our method on 6 MT systems that translate between English and German.

1 Introduction

The evaluation of machine translation (MT) systems has been a central research topic in recent years (cp. (Sparck-Jones and Galliers, 1995; King, 1996)). Many suggestions have focussed on measuring the translation quality (e.g. error classification in (Flanagan, 1994) or post editing time in (Minnis, 1994)). These measures are time-consuming and difficult to apply. But translation quality rests on the linguistic competence of the MT system which again is based first and foremost on grammatical coverage and lexicon size. Testing grammatical coverage can be done by using a test suite (cp. (Nerbonne et al., 1993; Volk, 1995)). Here we will advocate a probing method for determining the lexical coverage of commercial MT systems.

We have evaluated 6 MT systems which translate between English and German and which are all positioned in the low price market (under US$ 1500).

- German Assistant in Accent Duo V2.0 (developer: MicroTac/Globalink; distributor: Accent)
- Langenscheidts T1 Standard V3.0 (developer: GMS; distributor: Langenscheidt)
- Personal Translator plus V2.0 (developer: IBM; distributor: von Rheinbaben & Busch)
- Power Translator Professional (developer/distributor: Globalink)[1]
- Systran Professional for Windows (developer: Systran S.A.; distributor: Mysoft)
- Telegraph V1.0 (developer/distributor: Globalink)

The overall goal of our evaluation was a comparison of these systems resulting in recommendations on which system to apply for which purpose. The evaluation consisted of compiling a list of criteria for self evaluation and three experiments with external volunteers, mostly students from a local interpreter school. These experiments were performed to judge the information content of the translations, the translation quality, and the user-friendliness.

The list of criteria for self evaluation consisted of technical, linguistic and ergonomic issues. As part of the linguistic evaluation we wanted to determine the lexical coverage of the MT systems since only some of the systems provide figures on lexicon size in the documentation.

Many MT system evaluations in the past have been white-box evaluations performed by a testing team in cooperation with the developers (see (Falkedal, 1991) for a survey). But commercial MT systems can only be evaluated in a black-box setup since the developer typically will not make the source code and even less likely the linguistic source data (lexicon and grammar) available. Most of the evaluations described in the literature have centered around one MT system. But there are

[1] Recently a newer version has been announced as "Power Translator Pro 6.2".

hardly any reports on comparative evaluations. A noted exception is (Rinsche, 1993), which compares SYSTRAN[2], LOGOS and METAL for German - English translation[3]. She uses a test suite with 5000 words of authentic texts (from an introduction to Computer Science and from an official journal of the European Commission). The resulting translations are qualitatively evaluated for lexicon, syntax and semantics errors. The advantage of this approach is that words are evaluated in context. But the results of this study cannot be used for comparing the sizes of lexicons since the number of error tokens is given rather than the number of error types. Furthermore it is questionable if a running text of 5000 words says much about lexicon size, since most of this figure is usually taken up by frequent closed class words.

If we are mainly interested in lexicon size this method has additional drawbacks. First, it is time-consuming to find out if a word is translated correctly within running text. Second, it takes a lot of redundant translating to find missing lexical items.

So, if we want to compare the lexicon size of different MT systems, we have to find a way to determine the lexical coverage by executing the system with selected lexical items. We therefore propose to use a special word list with words in different frequency ranges to probe the lexicon efficiently.

2 Our method of probing the lexicon

Lexicon size is an important selling argument for print dictionaries and for MT systems. The counting methods however are not standardized and therefore the advertised numbers need to be taken with great care (for a discussion see (Landau, 1989)). In a similar manner the figures for lexicon size in MT systems ("a lexicon of more than 100.000 words", "more than 3.000 verbs") need to be critically examined. While we cannot determine the absolute lexicon size with a black-box test we can determine the relative lexical coverage of systems dealing with the same language pair.

When selecting the word lists for our lexicon evaluation we concentrated on adjectives, nouns, and verbs. We assume that the relatively small number of closed class words like determiners, pronouns, prepositions, conjunctions, and adverbs must be exhaustively included in the lexicon. For each of the three word classes in question (Adj, N, V) we tested words with high, medium, and low absolute frequency. We expected that words with high frequency should all be included in the lexicon, whereas words with medium and low frequency should give us a comparative measure of lexicon size. With these word lists we computed:

1. What percentage of the test words is translated?

2. What percentage of the test words is correctly translated?

The difference between 1. and 2. stems mostly from the fact that the MT systems regard unknown words as compounds, split them up into known units, and translate these units. Obviously this results in sometimes bizarre word creations (see section 2.3).

Our evaluation consisted of three steps. First, we prepared the word lists. Second, we ran the tests on all systems. Finally, we evaluated the output. These steps had to be done for both translation directions (German to English and vice versa), but here we concentrate on English to German.

2.1 Preparation of the word lists

We extracted the words for our test from the CELEX database. CELEX (Baayen, Piepenbrock, and van Rijn, 1995) is a lexical database for English, German and Dutch. It contains 51,728 stems for German (among them 9,855 adjectives; 30,715 nouns; 9,400 verbs) and 52,447 stems for English (among them 9,214 adjectives; 29,494 nouns; 8,504 verbs). This database also contains frequency data which for German were derived from the Mannheim corpus of the "Institut für deutsche Sprache" and for English were computed from the Cobuild corpus of the University of Birmingham. Looking at the frequency figures we decided to take:

- The 100 most frequent adjectives, nouns, verbs.

- 100 adjectives, nouns, verbs with frequency 25 or less. Frequency 25 was chosen because it is a medium frequency for all three word classes.

- The first 100 adjectives, nouns, verbs with frequency 1.[4]

[2]SYSTRAN is not to be confused with Systran Professional for Windows. SYSTRAN is a system with a development history dating back to the seventies. It is well known for its long-standing employment with the European Commission.

[3]Part of the study is also concerned with French - English translation.

[4]CELEX also contains entries with frequency 0, but we wanted to assure a minimal degree of commonness by selecting words with frequency 1. Still, many words with frequency 1 seem exotic or idiosyncratic uses.

Unfortunately the CELEX data contain some noise especially for the German entries. This meant that the extracted word lists had to be manually checked. One problem is that some stems occur twice in the list. This is the case if a verb is used with a prefix in both the separable and the fixed variant (as e.g. *übersetzen* engl. *to translate* vs. *to ferry across*). Since our test does not distinguish these variants we took only one of these stems. Another problem is that the frequency count is purely wordform-based. That means, if a word is frequently used as an adverb and seldom as a verb the count of the total number of occurrences will be attributed to both the adverb and the verb stem. Therefore, some words appear at strange frequency positions. For example the very unusual German verb *heuen* (engl. *to make hay*) is listed among the 100 most frequent verbs. This is due to the fact that its 3rd person past tense form is a homograph of the frequent adverb *heute* (engl. *today*). Such obviously misplaced words were eliminated from the list, which was refilled with subsequent items in order to contain exactly 100 words in each frequency class of each word.

The English data in CELEX are more reliable. The frequency count has been disambiguated for part of speech by manually checking 100 occurrences of each word-form and thus estimating the total distribution. In this way it has been determined that *bank* is used as a noun in 97% of all occurrences (in 3% it is a verb). This does not say anything about the distribution of the different noun readings (*financial institution* vs. *a slope alongside a river* etc.).

If a word is the same in English and in German (as e.g. *international, Squaw*) it must also be excluded from our test list. This is because some systems insert the source word into the target sentence if the source word (and its translation) is not in the lexicon. If source word and target word are identical we cannot determine if the word in the target sentence comes from the lexicon or is simply inserted because it is unknown.

After the word lists had been prepared, we constructed a simple sentence with every word since some systems cannot translate lists with single word units. With the sentence we were trying to get each system to translate a given word in the intended part of speech. For German we chose the sentence templates:

(1) Es ist ⟨adjective⟩.
Ein ⟨noun⟩ ist gut.
Wir müssen ⟨verb⟩.

Adjectives were tested in predicative use since this is the only position where they appear uninflected. Nouns were embedded within a simple copula sentence. The indefinite article for a noun sentence was manually adjusted to 'eine' for female gender nouns. Nouns that occur only in a plural form also need special treatment, i.e. a plural determiner and a plural copula form. Verbs come after the modal verb *müssen* because it requires an infinitive and it does not distinguish between separable prefix verbs and other verbs. On similar reasons we took for English:

(2) This is ⟨adjective⟩.
The ⟨noun⟩ can be nice.
We ⟨verb⟩.

The modal *can* was used in noun sentences to avoid number agreement problems for plural-only words like *people*. Our sentence list for English nouns thus looked like:

(3) 1. The time can be nice.
2. The man can be nice.
3. The people can be nice.
...
300. The unlikelihood can be nice.

2.2 Running the tests

The sentence lists for adjectives, nouns, and verbs were then loaded as source document in one MT system after the other. Each system translated the sentence lists and the target document was saved. Most systems allow to set a subject area parameter (for subjects such as finances, electrical engineering, or agriculture). This option is meant to disambiguate between different word senses. The German noun *Bank* is translated as English *bank* if the subject area is finances, otherwise it is translated as *bench*. No subject area lexicon was activated in our test runs. We concentrated on checking the general vocabulary.

In addition Systran allows for the selection of document types (such as prose, user manuals, correspondence, or parts lists). Unfortunately the documentation does not tell us about the effects of such a selection. No document type was selected for our tests.

Running the tests takes some time since 900 sentences need to be translated by 6 systems. On our 486-PC the systems differ greatly in speed. The fastest system processes at about 500 words per minute whereas the slowest system reaches only 50 words per minute.

2.3 Evaluating the tests

After all the systems had processed the sentence lists, the resulting documents were merged for ease

of inspection. Every source sentence was grouped together with all its translations. Example 4 shows the English adjective *hard* (frequency rank 41) with its translations.

(4)

41.		This is hard.
41.	G. Assistant	Dieser ist hart.
41.	Lang. T1	Dies ist schwierig.
41.	Personal Tr.	dies ist schwer.
41.	Power Tr.	Dieses ist hart.
41.	Systran	Dieses ist hart.
41.	Telegraph	Dies ist hart.

Note that the 6 MT systems give three different translations for *hard* all of which are correct given an appropriate context. It is also interesting to see that the demonstrative pronoun *this* is translated into different forms of its equivalent pronoun in German.

These sentence groups must then be checked manually to determine whether the given translation is correct. The translated sentences were annotated with one of the following tags:

u (unknown word) The source word is unknown and is inserted into the translation. Seldom: The source word is a compound, part of which is unknown and inserted into the translation (*the warm-heartedness : das warme heartedness*).

w (wrong translation) The source word is incorrectly translated either because of an incorrect segmentation of a compound (*spot-on : erkennen-auf/Stelle-auf* instead of *haargenau/exakt*) or (seldom) because of an incorrect lexicon entry (*would : würdeten* instead of *würden*).

m (missing word) The source word is not translated at all and is missing in the target sentence.

wf (wrong form) The source word was found in the lexicon, but it is translated in an inappropriate form (e.g. it was translated as a verb although it must be a noun) or at least in an unexpected form (e.g. it appears with duplicated parts (*windscreen-wiper : Windschutzscheibenscheibenwischer*)).

s (sense preservingly segmented) The source word was segmented and the units were translated. The translation is not correct but the meaning of the source word can be inferred (*unreasonableness : Vernunftlos-heit* instead of *Unvernunft*).

f (missing interfix (nouns only)) The source word was segmented into units and correctly translated. But the resulting German compound is missing an interfix (*windscreen-wiper : Windschutzscheibe-Wischer*).

wd (wrong determiner (nouns only)) The source word was correctly translated but comes with an incorrect determiner (*wristband : die Handgelenkband* instead of *das Handgelenkband*).

c (correct) The translation is correct.

Out of these tags only **u** can be inserted automatically when the target sentence word is identical with the source word. Some of the tested translation systems even mark an unknown word in the target sentence with special symbols. All other tags had to be manually inserted. Some of the low frequency items required extensive dictionary look-up to verify the decision. After all translations had been tagged, the tags were checked for consistency and automatically summed up.

3 Results of our evaluation

The MT systems under investigation translate between English and German and we employed our evaluation method for both translation directions. Here we will report on the results for translating from English to German. First, we will try to answer the question of what percentage of the test words was **translated at all** (correctly or incorrectly). This figure is obtained by taking the unknown words as negative counts and all others as positive counts. We thus obtained the triples in table 1. The first number in a triple is the percentage of positive counts in the high frequency class, the second number is the percentage of positive counts in the medium frequency class, and the third number is the percentage of positive counts in the low frequency class.

In table 1 we see immediately that there were no unknown words in the high frequency class for any of the systems. The figures for the medium and low frequency classes require a closer look. Let us explain what these figures mean, taking the German Assistant as an example: 14 adjectives (14 nouns, 21 verbs) of the medium frequency class were unknown, resulting in 86% adjectives (86% nouns, 79% verbs) getting a translation. In the low frequency class 49 adjectives, 53 nouns, and 61 verbs got a translation.

The average is computed as the mean value over the three word classes. Comparing the systems' averages we can observe that Personal Translator scores highest for all frequency classes. Langenscheidts T1 and Telegraph are second best with about the

	G. Assistant	Lang. T1	Personal Tr.	Power Tr.	Systran	Telegraph
adjectives	100/86/49	100/98/66	100/95/84	100/87/54	100/49/31	100/97/59
nouns	100/86/53	100/91/62	100/97/78	100/83/53	100/59/32	100/94/63
verbs	100/79/61	100/97/73	100/97/88	100/84/55	100/61/37	100/93/75
average	100/84/54	100/95/67	100/96/83	100/85/54	100/56/33	100/95/66

Table 1: Percentage of words translated correctly or incorrectly

	G. Assistant	Lang. T1	Personal Tr.	Power Tr.	Systran	Telegraph
adjectives	100/79/24	100/92/36	100/94/77	100/86/49	100/47/23	100/96/53
nouns	99/83/38	100/88/50	100/95/74	100/81/47	100/57/27	100/92/53
verbs	97/78/50	99/93/59	100/97/86	100/84/50	100/61/33	100/93/73
average	99/80/37	100/91/48	100/95/79	100/84/49	100/55/28	100/94/60

Table 2: Percentage of correctly translated words

same scores. German Assistant and Power Translator rank third while Systran clearly has the lowest scores. This picture becomes more detailed when we look at the second question.

The second question is about the percentage of the test words that are **correctly translated**. For this, we took unknown words, wrong translations, and missing words as negative counts and all others as positive counts. Note that our judgement does not say that a word is translated correctly in a given context. It merely states that a word is translated in a way that is understandable in some context.

Table 2 gives additional evidence that Personal Translator has the most elaborate lexicon for English to German translation while German Assistant and Systran have the least elaborate. Telegraph is on second position followed by Langenscheidts T1 and Power Translator. We can also observe that there are only small differences between the figures in table 1 and table 2 as far as the high and medium frequency classes are concerned. But there are differences of up to 30% for the low frequency class. This means that we will get many wrong translations if a word is not included in the lexicon and has to be segmented for translation.

While annotating sentences with the tags we observed that verbs obtained many 'wrong form' judgements (20% and more for the low frequency class). This is probably due to the fact that many English verbs in the low frequency class are rare uses of homograph nouns (e.g. *to keyboard, to pitchfork, to section*). If we omit the 'wrong form' tags from the positive count (i.e. we accept only words that are correct, sense preservingly segmented, or close to correct because of minor orthographical mistakes) we obtain the figures in table 3.

In this table we can see even clearer the wide coverage of the Personal Translator lexicon because the system correctly recognizes around 70% of all low frequency words while all the other systems figure around 40% or less. It is also noteworthy that the Systran results differ only slightly between table 2 and table 3. This is due to the fact that Systran does not give many wrong form (wf) translations. Systran does not offer a translation of a word if it is in the lexicon with an inappropriate part of speech. So, if we try to translate the sentence in example 5 Systran will not offer a translation although *keyboard* as a noun is in the lexicon. All the other systems give the noun reading in such cases.

(5) We keyboard.

So the difference between the figures in tables 2 and 3 gives an indication of the precision that we can expect when the translation system deals with infrequent words. The smaller the difference, the more often the system will provide the correct part of speech (if it translates at all).

3.1 Some observations

NLP systems can widen the coverage of their lexicon considerably if they employ word-building processes like composition and derivation. Especially derivation seems a useful module for MT systems since the meaning shift in derivation is relatively predictable and therefore the derivation process can be recreated in the target language in most cases.

It is therefore surprising to note that all systems in our test seem to lack an elaborate derivation module. All of them know the noun *weapon* but none is able to translate *weaponless*, although the English derivation suffix *-less* has an equivalent in German

	G. Assistant	Lang. T1	Personal Tr.	Power Tr.	Systran	Telegraph
adjectives	90/72/21	97/74/28	99/92/69	92/75/43	97/43/21	92/84/44
nouns	98/80/30	100/83/44	100/94/73	98/77/44	100/55/24	99/90/46
verbs	97/63/16	97/85/26	99/91/67	100/76/22	100/53/13	99/86/41
average	95/72/22	98/81/33	99/92/70	97/76/36	99/50/19	97/87/44

Table 3: Percentage of correctly translated words (without 'wrong forms')

	G. Assistant	Lang. T1	Personal Tr.	Power Tr.	Systran	Telegraph
wd-nouns	8	2	-	7	0	2

Table 4: Number of incorrect gender assignments

-los. German Assistant treats this word as a compound and incorrectly translates it as *Waffe-weniger* (engl. *less weapon*). Due to the lack of derivation modules, words like *uneventful, unplayable, tearless, or thievish* are either in the lexicon or they are not translated. Traces of a derivational process based on prefixes have been found for Langenscheidts T1 and for Personal Translator. They use the derivational prefix *re-* to translate English *reorient* as German *orientieren wieder* which is not correct but can be regarded as sense preserving.

On the other hand all systems employ segmentation on unknown compounds. Example 6 shows the different translations for a compound noun. The marker 'M' in the Langenscheidts T1 translation indicates that the translation has been found via compound segmentation. While *Springpferd, Turnpferd* or simply *Pferd* could count as correct translations of *vaulting-horse, Springen-Pferd* can still be regarded as sense-preservingly segmented.

(6)

English:	vaulting-horse	
G. Assistant	Gewölbe-Pferd	**w**
Lang. T1	⟨M[Springpferd]⟩	**c**
Personal Tr.	Wölbungspferd	**w**
Power Tr.	Springen - Pferd	**s**
Systran	Vaultingpferd	**u**
Telegraph	Gewölbe-Kavallerie	**w**

An example of a verb compound that gets a translation via segmentation is *to tap-dance* and an adjective compound example is *sweet-scented*. All of these examples are hyphenated compounds. If we look at compounds that form an orthographic unit like *vestryman, waterbird* we can only find evidence for segmentations by Langenscheidts T1 and German Assistant. These findings only relate to translating from English to German. Working in the opposite direction all systems perform segmentation of orthographic unit compounds since this is a very common feature of German.

As another side effect we used the lexicon evaluation to check for agreement within the noun phrase. Translating from English to German the MT system has to get the gender of the German noun from the lexicon since it cannot be derived from the English source. We can check if these nouns get the correct gender assignment if we look at the form of the determiner. Table 4 gives the number of incorrect determiner selections (over all frequency classes).

Since gender assignment in choosing the determiner is such a basic operation all systems are able to do this in most cases. But in particular if noun compounds are segmented and the translation is synthesized this operation sometimes fails. Personal Translator does not give a determiner form in these cases. It simply gives the letter 'd' as the beginning letter of all three different forms (*der, die, das*).

3.2 Comparing translation directions

Comparing the results for English to German translation with German to English is difficult because of the different corpora used for the CELEX frequencies. Especially it is not evident whether our medium frequency (25 occurrences) leads to words of similar prominence in both languages. Nevertheless our results indicate that some systems focus on either of the two translation directions and therefore have a more elaborate lexicon in one direction. This can be concluded since these systems show bigger differences than the others. For instance, Telegraph, Systran and Langenscheidts T1 score much better for German to English. For Telegraph the rate of unknown words dropped by 2% for medium frequency and by 12% for low frequency, for Systran the same rate dropped by 36% for medium frequency and by 33% for low frequency words, and for Langenscheidts T1 the rate dropped by 1% for medium frequency and by 16% for low frequency. The latter

reflects the figures in the Langenscheidts T1 manual, where they report an inbalance in the lexicon of 230'000 entries for German to English and 90'000 entries for the opposite direction. Personal Translator again ranks among the systems with the widest coverage while German Assistant shows the smallest coverage.

4 Conclusions

As more translation systems become available there is an increasing demand for comparative evaluations. The method for checking lexical coverage as introduced in this paper is one step in this direction. Taking the most frequent adjectives, nouns, and verbs is not very informative and mostly serves to anchor the method. But medium and low frequency words give a clear indication of the underlying relative lexicon size. Of course, the introduced method cannot claim that the relative lexicon sizes correspond exactly to the computed percentages. For this the test sample is too small. The method provides a plausible hypothesis but it cannot prove in a strict sense that one lexicon necessarily is bigger than another. A proof, however, cannot be expected from any black-box testing method.

We mentioned above that some systems subclassify their lexical entries according to subject areas. They do this to a different extent.

Langenscheidts T1 has a total of 55 subject areas. They are sorted in a hierarchy which is three levels deep. An example is Technology with its subfields Space Technology, Food Technoloy, Technical Norms etc. Multiple subject areas from different levels can be selected and prioritized.

Personal Translator has 22 subject areas. They are all on the same level. Examples are: Biology, Computers, Law, Cooking. Multiple selections can be made, but they cannot be prioritized.

Power Translator and Telegraph do not come with built-in subject dictionaries but these can be purchased separately and added to the system.

Systran has 22 "Topical Glossaries", all on the same level. Examples are: Automotive, Aviation/Space, Chemistry. Multiple subject areas can be selected and prioritized.

Our tests were run without any selection of a subject area. We tried to check if a lexicon entry that is marked with a subject area will still be found if no subject area is selected. This check can only be performed reliably for Langenscheidt T1 since this is the only system that makes the lexicon transparent to the user to the point that one can access the subject area of every entry. Personal Translator only allows to look at an entry and its translation options, but not at its subject marker, and Systran does not allow any access to the built-in lexicon. For Langenscheidts T1 we tested the word *compiler* which is marked with *data processing* and *computer software*. This lexical entry does not have any reading without a subject area marker, but the word is still found at translation if no subject area is chosen. That means that a subject area, if chosen, is used as disambiguator, but if translating without a subject area the system has access to the complete lexicon.

In this respect our tests have put Power Translator and Telegraph at a disadvantage since we did not extend their lexicons with any add-on lexicons. Only their built-in lexicons were evaluated here.

Of course, lexical coverage by itself does not guarantee a good translation. It is a necessary but not a sufficient condition. It must be complemented with lexical depth and grammatical coverage. Lexical depth can be evaluated in two dimensions. The first dimension describes the number of readings available for an entry. A look at some common nouns that received different translations from our test systems reveals that there are big differences in this dimension which are not reflected by our test results. Table 7 gives the number of readings for the word *order* ('N' standing for noun readings, 'V' for verbal, 'Prep' for prepositional, and 'Phr' for phrasal readings).

(7)

G. Assistant	9 N	3 V	1 Prep	
Lang. T1	4 N	4 V		
Personal Tr.	6 N	5 V		
Power Tr.	1 N	1 V	1 Prep	
Systran	n.a.			
Telegraph	10 N	4 V		2 Phr

There is no information for Systran since the built-in lexicon cannot be accessed. German Assistant contains a wide variety of readings although it scored badly in our tests. Power Translator on the contrary gives only the most likely readings. Still, there remains the question of whether a system is able to pick the most appropriate reading in a given context, which brings us to the second dimension.

The second dimension of lexical depth is about the amount of syntactic and semantic knowledge attributed to every reading. This also varies a great deal. Telegraph offers 16 semantic features (ani-

mate, time, place etc.), German Assistant 9 and Langenscheidts T1 5. Power Translator offers few semantic features for verbs (movement, direction). The fact that these features are available does not entail that they are consistenly set at every appropriate reading. And even if they are set, it does not follow that they are all optimally used during the translation process.

To check these lexicon dimensions new tests need to be developped. We think that it is especially tricky to get to all the readings along the first dimension. One idea is to use the example sentences listed with the different readings in a comprehensive print dictionary. If these sentences are carefully designed they should guide an MT system to the respective translation alternatives.

Our method for determining lexical coverage could be refined by looking at more frequency classes (e.g. an additional class between medium and low frequency). But since the results of working with one medium and one low frequency class show clear distinctions between the systems, it is doubtful that the additional cost of taking more classes will provide significantly better figures.

The method as introduced in this paper requires extensive manual labor in checking the translation results. Carefully going through 900 words each for 6 systems including dictionary look-up for unclear cases takes about 2 days time. This could be reduced by automatically accessing translation lists or reliable bilingual dictionaries. Judging sense-preserving segmentations or other close to correct translations must be left over to the human expert.

A special purpose translation list could be incrementally built up in the following manner. For the first system all 900 words will be manually checked. All translations with their tags will be entered into the translation list. For the second system only those words will be checked where the translation differs from the translation saved in the translation list. Every new judgement will be added to the translation list for comparison with the next system's translations.

5 Acknowledgements

I would like to thank Dominic A. Merz for his help in performing the evaluation and for many helpful suggestions on earlier versions of the paper.

References

Baayen, R. H., R. Piepenbrock, and H. van Rijn. 1995. The CELEX lexical database (CD-ROM). Linguistic Data Consortium, University of Pennsylvania.

Falkedal, Kirsten. 1991. Evaluation Methods for Machine Translation Systems. An historical overview and a critical account. ISSCO. University of Geneva. Draft Report.

Flanagan, Mary A. 1994. Error classification for MT evaluation. In *Technology partnerships for crossing the language barrier: Proceedings of the 1st Conference of the Association for Machine Translation in the Americas*, pages 65–71, Washington,DC. Association for Machine Translation in the Americas.

Landau, Sidney I. 1989. *Dictionaries. The art and craft of lexicography.* Cambridge University Press, Cambridge. first published 1984.

King, Margaret. 1996. Evaluating natural language processing systems. *CACM*, 39(1):73–79.

Minnis, Stephen. 1994. A simple and practical method for evaluating machine translation quality. *Machine Translation*, 9(2):133–149.

Rinsche, Adriane. 1993. *Evaluationsverfahren für maschinelle Übersetzungssysteme - zur Methodik und experimentellen Praxis.* Kommission der Europäischen Gemeinschaften, Generaldirektion XIII; Informationstechnologien, Informationsindustrie und Telekommunikation, Luxemburg.

Nerbonne, J., K. Netter, A.K. Diagne, L. Dickmann, and J. Klein. 1993. A diagnostic tool for german syntax. *Machine Translation (Special Issue on Evaluation of MT Systems), (also as DFKI Research Report RR-91-18)*, 8(1-2):85–108.

Sparck-Jones, K. and J.R. Galliers. 1995. *Evaluating Natural Language Processing Systems. An Analysis and Review.* Number 1083 in Lecture Notes in Artificial Intelligence. Springer Verlag, Berlin.

Volk, Martin. 1995. *Einsatz einer Testsatzsammlung im Grammar Engineering*, volume 30 of *Sprache und Information.* Niemeyer Verlag, Tübingen.