The PLUG project: Parallel corpora in Linköping, Uppsala, Göteborg: aims and achievements

Anna Sågvall Hein

Department of Linguistics, Uppsala University

Abstract

In this paper we present the aims and achievements of the PLUG project. It is a cooperative Swedish project focusing on the generation of translation data from sentence-aligned bitext with Swedish as the source or the target. A sentence-aligned quadrilingual corpus was established and used as a testbed. Two systems for word linking and contrastive lexical extraction were evaluated and improved with the aim of combining them into a common system. The common system will run as an application of a modular corpus tool also created in the project. The basic principles of the word linking systems are outlined and illustrative results are presented and discussed with regard to recall, precision, and application in example-based machine translation, enhanced machine translation, transfer-based machine translation and human translation. Further processing for these applications as well as integration issues remain to be explored. Finally, the extraction of syntactic translation data is an issue that remains to be approached. Focus will be set on verb valency with imperative and infinitive clauses as basic frames.

1. Introduction

The Plug project is a Swedish co-operative project aimed at the development, evaluation and application of programs for alignment and data generation from parallel corpora with Swedish as the source or the target language. Applications include machine translation, human translation, computer-aided translation, translation data-bases, translation dictionaries, and translator's training. The participating parties are the Department of Computer and Information Science at Linköping University, the Department of Linguistics at Uppsala University, and the Department of Swedish Language at Göteborg University (980401–981231) and Anna Sågvall Hein is the project leader. The project is funded jointly by the Swedish Research Council for the Humanities and Social Sciences, HSFR, and the Board for Technical and Industrial Development, NUTEK, in the framework of the Swedish Language Technology Programme for Research and Development. It will run during the period 1998-04-01 – 2000-03-31, see further http://stp.ling.uu.se/~corpora/plug/.

An important aspect of the project is to increase co-operation and co-ordination between different groups in Sweden that work on parallel corpora and their applications. A first result of this co-operation is an overview of state-of-the-art and existing Swedish resources that was compiled as part of the pilot phase of the project (Ahrenberg *et al*. 1998). Another goal is to raise the awareness in the

commercial sector for the application of translation support in a production environment. A step in this direction was taken with the study of machine translation with special reference to Swedish and Sweden that was initiated by the Swedish Department of Industry to which representatives of the Plug project contributed (se further NUTEK 1999).

Here we report on the aims of the project and the achievements made so far.

2. Background

The use of computers to support translation and translation studies is steadily on the increase. For certain text types machine translation has proven useful, while translator's workbenches are rapidly becoming common tools for many translation companies and freelance translators.

Moreover, a lot of the infrastructure needed in translation projects, such as terminology work and editing can also be supported by computer aids. It goes without saying that such tools are more useful the more support they can give for the languages they are supposed to be applied to.

The use of corpus-based translation data is in the mainstream of research and development of machine translation systems today. An example of a memory-based, multilingual translation system is ESTeam BTR (NUTEK 1999:23). It is the only more advanced machine translation system on the market today including Swedish. The translation memory of BTR contains full sentences as well as sub-sentence segments of various kinds. BTR aims at achieving acceptable translations of simple, repetitive text. A successful example of fully automatic translation with BTR is the translation of trademark data.

An example of one of the major vendors making use of corpus-based translation data to a larger extent is Systran. The improvement of the translation quality that may result from the enhancement with corpus data is demonstrated for the translation of medical text from French to English (http://onlinetrans.com/Webling.html). Systran has expressed an interest in incorporating Swedish among the languages that it supports.

Another candidate for a Swedish machine translation system is the Multra prototype in need for a large-scale, corpus-based translation data (Sågvall Hein 1994, 1997).

The use of previous translations is also the core idea behind the translator's workbenches. The translation memories used by these workbenches typically comprise sentence-aligned bitext. However, when it comes to example-based machine translation systems, sub-sentence segments down to the single word have proven to be useful. We are only in the beginning of investigating the various ways in which sentence bitext may be segmented into useful segments for

machine translation purposes. A challenge for the future is the integration of example-based machine translation and earlier approaches based on transfer translation.

3. Aims and results

A long-term goal of the project is to explore how Swedish translation data of various kinds can be extracted from bitext, formalised and fed into existing machine translation systems thereby making them work as proper tools. Another goal is to provide translators, who translate to or from Swedish, with useful support tools such as translation memories and corpus-based translation dictionaries. Actually, we don't restrict the use of support tools to the translation process itself, but to all phases of the process, including terminology work, preparation and validation of translation data-bases, creation of multilingual dictionaries, editing, and language learning. Methods for data extraction from parallel corpora provide a base technology for the development and enhancement of translation tools for humans as well as for machines.

The project considers the automatic extraction of translation data of various kinds, including general vocabulary, terminology, phraseology, valency information, and sentence patterns. It has provided concrete results in terms of

- a quadrilingual sentence-aligned corpus with Swedish either as the source language or the target language
- methods for searching the corpus
- two systems for word alignment and contrastive lexical data extraction
- Uplug - a modular corpus tool
- data from experiments with word and phrase linking and contrastive lexical data extraction
- an implemented approach to the evaluation of automatically generated word and phrase links

It will provide concrete results in terms of

- a database of translation data for English ←→ Swedish and German ←→ Swedish with a graphical search interface
- transfer rules for transfer based machine translation (Multra)
- translation rules for enhanced machine translation (Systran)
- illustrative rules for example based machine translation

4. A quadrilingual sentence-aligned corpus

A common project corpus of four languages (Swedish, English, German, Italian) representing three genres (technical text, political text, and literary text) was established. It includes contributions of parallel text from the three participating departments. By a parallell text we understand a source text with a translation into a target language. Text files for the corpus were delivered in different formats (Ahrenberg *et al*. 1999). Before being included in the common corpus, they were all encoded in XML using the plugXML.dtd (Tiedemann 1998b). Further, the texts were automatically aligned at the level of the sentence. For this purpose, the alignment algorithm of Gale and Church (1993) was used. Prior to the alignment, the texts were tokenized and split into technical sentences, i.e. orthographic sentences, head lines, list elements, and table cells. For a full account of the technical details of the project corpus, see Tiedemann 1998b.

Table 1: The complete Plug corpus

Language pair	*Number of words*
Swedish/English	1,169,165
Swedish/German	525,278
Swedish/Italian	493,636
Total	2,188,079

4.1 Technical text

Technical documentation represents the largest component of the corpus. It is more than three times as big as each of the other two components with a total size of about 1.35 million words. It origins from two different sources which were contributed by Linköping University and Uppsala University. All the three language pairs are included: Swedish/English, Swedish/German, and Swedish/ Italian.

The contribution of Uppsala University consists of parts of the Scania 1995 Corpus. The original texts were provided by Scania CV AB for an earlier study aiming at the establishment of a controlled vocabulary for truck and bus maintenance (Almqvist and Sågvall Hein 1996; Sågvall Hein 1997; Sågvall Hein *et al*. 1997).

Table 2: The Scania corpus

Language pair	*Number of words*
Swedish -> English	385,289
Swedish -> German	337,188
Swedish -> Italian	343,129

The Swedish source version of Scania 1995 amounts to 172,259 words. The English, German and Italian versions are direct translations of the Swedish original. Linköping contributed manuals for Microsoft's software packages MS Excel and MS Access. In both cases English is the source language.

Table 3: The Microsoft corpus

Language pair	*Text*	*Number of words*
English -> Swedish	MS Excel	124,961
English -> Swedish	MS Access	163,173

4.2 Political and administrative Text

The political and administrative component of the corpus includes contributions from Göteborg and Uppsala. The main part represents Swedish/English and Swedish/German texts. There is also a minor part of Italian/Swedish bitext. The component of political texts with a total size of about 410,000 words is relatively small compared to the technical documentation component. However, the size is comparable to that of the component of literary texts which will be described in the next section. The main part consists of administrative texts from the European Union which were collected and aligned at the Department of Swedish Language in Göteborg. They are part of the PEDANT corpus (Ridings 1998). Eventhough the translation history of these texts is not quite clear, there is no doubt that Swedish is the target.

Table 4: Texts from the European Union

Language pair	*Number of words*
English -> Swedish	186,111
German -> Swedish	180,312
Italian -> Swedish	28,196

The declarations of the Swedish government contributed by Uppsala University represent the smallest portion of the corpus. The texts were translated from Swedish to English and German, respectively.

Table 5: Declarations of policy of the Swedish Government

Language pair	*Number of words*
Swedish -> English	8,011
Swedish -> German	7,778

4.3 Literary text

The literary component of the project corpus includes contributions from Linköping university and Göteborg University. The total size of 423,931 words

is comparable to the political text component. Linköping's contribution includes two novels which were translated from English to Swedish and Göteborg's contribution contains translations of two Swedish novels to Italian.

Linköping University provided two English/Swedish bitexts which origin from the novels "A Guest of Honour" by Nadine Gordimer and "To Jerusalem and back: a personal account" by Saul Bellow. The texts were originally provided by the Swedish Language Bank in Göteborg.

Table 6: English/Swedish novels

Language pair	*Text*	*Number of words*
English -> Swedish	A Guest of Honour.	169,554
English -> Swedish	To Jerusalem and back: a personal acoount.	132,066

Göteborg provided two novels by Lars Gustafsson ("En kakelsättares eftermiddag" and "En biodlares död") which were translated from Swedish to Italian.

Table 7: Swedish/Italian novels

Language pair	*Text*	*Number of words*
Swedish -> Italian	En kakelsättares eftermiddag.	66,429
Swedish -> Italian	En biodlares död.	55,882

5. Searching the corpus

Various interfaces for searching the corpus were implemented. Below we give an example of how the corpus may be searched via the web:

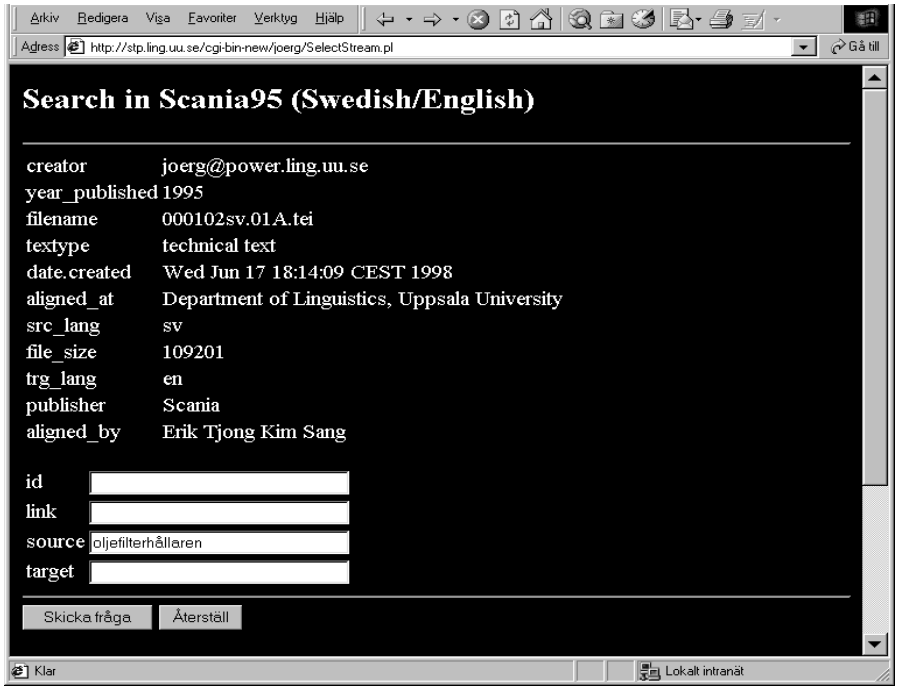

Table 8: Examples of link units

id	link	source	target
sventscan3888	1-1	I **oljefilterhållaren** sitter en överströmningsventil.	The oil filter retainer has an overflow valve.
sventscan3200	**2-1**	Undvik hudkontakt med kylvätska. Hudkontakt kan medföra irritation.	Avoid contact with the skin as this may cause irritation.
sventscan783	**1-2**	Skruvarna sträcks vid varje åtdragning, därför får skruvarna i en del förband återanvändas endast ett visst antal gånger.	Bolts are stretched each time they are tightened. For this reason, the bolts in some joints should only be reused a certain number of times.

The link units in the corpus are described with regard to origin (id), type of link relation, source sentence, and target sentence. All the four fields are searchable. The search key appears in bold face in the retrieved unit. The first example in Table 8 was retrieved via the search key **ojlefilterhållaren,** whereas the link type was the search key in the two following examples.

6. Two systems for word alignment

Linköping and Uppsala contributed their own systems for word and phrase alignment, Linköping Word Aligner (Ahrenberg *et al*. 1998) and Uppsala Word Aligner (Tiedmann 1998a), respectively. The two systems are fairly language independent, at least as regards Western European languages, and rely heavily on empirical data and statistical criteria (cf. Smadja *et al*. 1996; Melamed 1997b; Fung and Church 1994). In the course of the project, the two systems have been evaluated and improved. A statistically based system can never produce results that are a hundred per cent correct, and evaluation emerges as a core issue.

As in predecessor systems (see above) it is measured in terms of recall and precision, and a combination of them. By recall we understand the number of possible links that are retrieved, and by precision the accurateness of the link relations, see further Merkel and Ahrenberg (1999). An evalutation strategy making use of a gold standard has been implemented (Merkel and Ahrenberg, this volume) and applied to the two systems (Ahrenberg *et al*. forthcoming).

In the final end Linköping Word Aligner, LWA and Uppsala Word Aligner, UWA will be combined into one system, the Plug Word Aligner. It will run as an application of the Uplug-system, a modular corpus tool for parallel corpora being developed in the project (Tiedemann this volume). For an illustration of the basic principles of word alignment and contrastive lexical data extraction in the PLUG project we will use UWA.

6.1 Basic operation of Uppsala Word Aligner

In Fig.1 we present the basic operation of Uppsala Word Aligner, UWA. The parallel corpus that goes into the system is assumed to be a sentence aligned bitext. The preprocessing phase focuses on the segmentation of the text into link segments. A link segment is a single word token or a multi-word token, a phrase. Phrases are recognised via reference to an external dictionary (Wikholm *et al*. 1996) or generated from the text. Candidates for text generated phrases are recurrent sequences of tokens of a certain frequency and of a certain length. There are also some restrictions with regard to functional words and signs of punctuation at the beginning and the end of the phrase candidates. The identification of pairs of link segments, translation equivalents, is based on four main principles:

- Iterative size reduction - Link the safe cases first, remove them, and proceed with the rest. An example of an initial safe case is a sentence link unit, where the source or the target contains one single word, e.g. a head-line, a list item, or a table cell. Safe cases may also be provided by an external translation dictionary.
- String similarity evaluation. The assumption behind this criterion is that similar words are likely to be translation equivalents. An

extreme case of string similarity is string identity represented by proper names, acronyms etc. Simple string comparing algorithms based on character matching are used to measure the similarity between non-identical word pairs. Evaluations by threshold filtering produced a set of cognate pairs with reasonable precision for the considered language pairs. Further work in this direction has been carried out (Tiedemann 1999), but so far not been included in the system.

- Co-occurrence evaluation. The assumption behind this criterion is that translation equivalents tend to occur with the roughly the same frequency in roughly the same contexts. The implementation of this assumption is based on statistical measures to identify pairs with a high co-occurrence ratio. The Dice coefficient (Smadja 1996) was used to value pairs of link segments that were compiled from bilingual alignments.
- Evaluations of low frequent words. The criterion is based on the assumption that low frequent text units are translated into low frequent text. For this purpose, high and medium frequent words were removed from the alignments and remaining data were analysed for the retrieval of corresponding low frequent translation equivalents.

The post-processing component is used for filtering out inappropriate candidates.

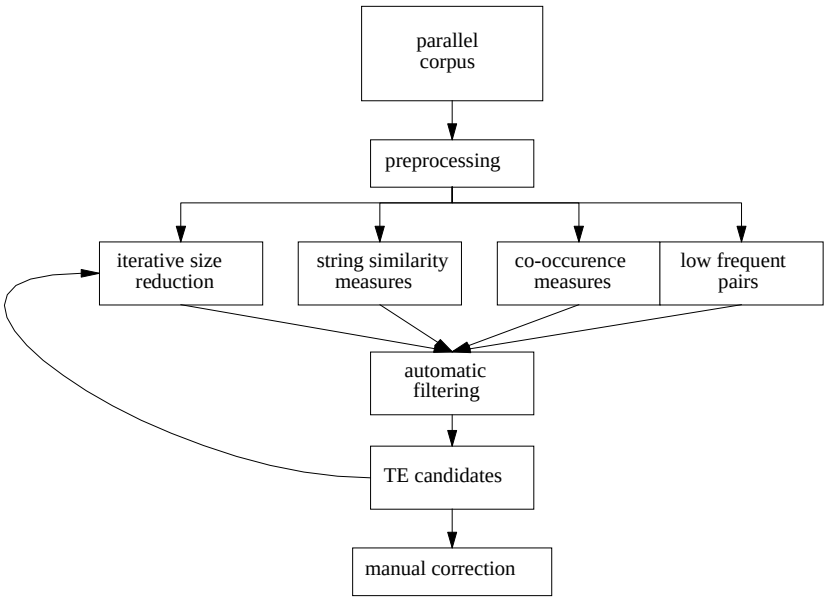

Figure 1: Extraction of translation equivalents from parallel corpora. Figure by Jörg Tiedemann, Department of Linguistics, Uppsala University.

6.2 Illustrative results

UWA as presented in Fig. 1 generates a set of translation equivalent candidates. The actual linking of their individual instances in the bitext is performed in a subsequent step not illustrated in Fig.1. The set of translation equivalent candidates provides raw data for building dictionaries and collecting data for example-based translation. Via the links to the bitext, contexts of any size within the limits of the bitext may be provided to illustrate how the words are used in the text.

For an illustration of the kinds of results achieved by UWA we present a fragment of the set of translation equivalent candidates that were generated when the system was applied to a Swedish-English bitext (Table 10). The bitext is part of the extended Plug corpus, including Scania 1998. No filtering was performed. We will look at the good cases and also at some of the shortcomings with regard to recall and precision and discuss how data of this kind may be further processed

Table 10: A fragment of a set of translation equivalent candidates generated by UWA

tryck ihop	compress
tryck ner bromspedalen	depress the brake pedal
tryck på	press
trycka	press, pressing
tryckas	pressed
tryckbegränsningsventil	pressure limiting valve
tryckbegränsningsventilen	pressure limiting valve
tryckbegränsningsventilens	pressure limiting valve
tryckbortfall	loss of pressure
tryckbricka	thrust washer
tryckbrickan	thrust washer
trycken	pressure, pressures
trycker	forces, press, pressed, presses, pushes, truck
trycket	pressure
tryckfall	pressure drop
tryckfallet	pressure drop
tryckfjäder	compression spring, spring
tryckfjädern	compression spring, spring
tryckfjäderns	compression spring
tryckfjädrar	compression springs, springs
tryckfjädrarna	compression springs, springs
tryckgivare	pressure sensor, pressure sensor/switch, pressure sensors
tryckgivare/vakt	pressure sensor/monitor

Several types of appropriate and interesting equivalence relations are spotted (Table 11).

Table 11: Examples of link relations generated by UWA

Swedish segment	Example	English segment	Example
one word unit	trycker	one word unit	presses
compound	tryckbricka	two word unit	thrust washer
compound	tryckbortfall	three word unit	loss of pressure
compound	tryckluftkompressor-kugghjul	four word unit	air compressor gear wheel
two word unit	tryckgivare/vakt	three word unit	pressure sensor/monitor
two word unit	tryck ihop	one word unit	compress
three word unit	tryck ner bromspedalen	four word unit	depress the brake pedal

As illustrated in Table 11, UWA finds translation relations holding between larger segments than the word, e.g. “Tryck ner bromspedalen” – “Depress the brake pedal”. Equivalents of this kind make example-based machine translation highly effective.

Basically, recall may not be measured from a study of retrieved relations only. We need a facit, a gold standard, to tell how many missing links there are. However, *alternative missing links* may be spotted in a close examination of the retrieved links. For instance, in Table 10 we find a suggestion for a translation relation between ”tryck på” and ”press”. It indicates an interpretation of the Swedish expression as a phrasal verb with an English single verb counterpart. It seems quite appropriate. We know, however, that there is a potential for an ambiguity relation in Swedish between phrasal verbs and nouns followed by prepositions. It may manifest itself as soon as there is an ambiguity between a single verb and a noun, and an adverb and a preposition. This is the case for “tryck på”, and to examine if this is so in the particular corpus, we search it. The search shows (see App.) that the string appears 14 times; 4 times as the phrasal verb, 5 times as the noun followed by a preposition (En. ”pressure of”), and 5 times as the final substring of two Swedish compounds (”resttryck”, ”öppningstryck”) followed by prepositions. The missing link representing the nominal relation is due to the English phrase generator. It fails to recognise ”pressure of” as a phrase. This is a kind of shortcoming that may easily be remedied in the further development of the system.

The retrieved relations present an appropriate basis for examining precision. Typically, morphology is not invariant during translation, as illustrated by several cases in Table 11. For instance, "tryckbegränsningsventil" appearing in three different Swedish forms (basic form; singular, definitie form, basic case; singular, definite form, genitive case) has a single English counterpart "pressure limiting valve". The problem may be approached in different ways depending on the intended application (see below).

Another precision problem concerns the phrases. Not seldom does the system find only a partial link, a part of an analytic, loose compound corresponding to a Swedish syntetic compound. An example of this is presented in Table 12.

Table 12: Examples of partial links

Swedish segment	*First English segment*	*Alternative English segment*
”tryckfjäder”	”compression spring”	”spring”
”tryckfjädern”	”compression spring”	”springs”
”tryckfjädrerns”	”compression spring”	
”tryckfjädrar”	”compression springs”	”springs”
”tryckfjädrarna”	”compression springs”	”springs”

For five of the inflectional forms of the Swedish compound ”tryckfjäder” two alternative translations are suggested, one appropriate, and one corresponding to a partial link. In cases like this one, the partial link may be filtered out via the subsumption relation. The subsumed, more specific link is kept, whereas the subsuming, more general one is removed.

6.3 Applications and further processing

The basic kinds of applications that we aim for in the PLUG project are:

1. collecting translation data for example-based machine translation
2. building dictionaries for enhanced direct machine translation
3. building dictionaries for transfer-based machine translation
4. building dictionaries for human translation

The first two applications (1.– 2.) use a direct translation strategy with no intermediary representation. This implies that the word linking system should be pushed as far as possible as regards precision. In the morphology example used above (“tryckbegränsningsventil”) maximal precision would mean finding the following links:

Table 13: Examples of Swedish – English word translations

Swedish	*English*
”tryckbegränsningsventil”	”pressure limiting valve”
”tryckbegränsningsventilen”	”the pressure limiting valve”
”tryckbegränsningsventilens”	”of the pressure limiting valve”

Pursuing the knowledge-lite strategy of UWA, this primarily means developing the English phrase generator further. The prospects for doing so are quite good.

An alternative to this line of development is a more knowledge intense approach. Basically, it amounts to translating the Swedish counterparts of the contrastive lexical data that were generated, using the transfer-based Multra system. A prototype version of the system is available, and what is needed for the implementation of this approach are the lexical correspondences that are generated by UWA. As a result, the kinds of translations illustrated in Table 12 will be generated.

In conclusion, if enhanced or example-based translation is the primary application of the word linking system, precision should be further improved. There are two strategies for doing so, either pushing UWA further by developing the English phrase generator, or actually translating the Swedish counterparts making use of the lexical correspondences generated by UWA. Certainly, a mixed approach may also be thought of.

Aiming for transfer-based machine translation, the second approach has to be followed. It represents one step towards the goal. A combination of example-based machine translation and transfer-based machine translation emerges as the most powerful setting for machine translation and one of the ultimate goals of a follow-up project of PLUG.

When it comes to building dictionaries for human translation, two important points should be made:

- The linking of all the instances of the contrastive lexical data that are generated makes it possible to connect the dictionary entries with contexts of any size within the limits of the actual text corpus.
- Among the contrastive lexical data generated by the system there are some that won't be found in a traditional dictionary and some that will be. Filtering out the corpus-specific words and phrases from the general vocabulary data is on important aspect.
- Lemmatising the lexical data and building a dictionary of lemmas rather than of word types seems to be called.

7. Conclusions and prospects for the future

The achievements made so far in the PLUG project provide useful lexical data for various tasks related to translation e.g. the building of dictionaries for enhanced direct machine translation, transfer-based machine translation, and human translation, and the collection of translation data for example-based machine translation.

How valuable the data prove to be depends on how far we may push recall and precision of the two systems, and the combined system. Continued work with that aim is in progress. A remaining issue is the elaboration, adaptation and actual

integration of the retrieved bilingual lexical data into the various translation applications aimed at. This issue includes the creation of a quadrilingual database with a graphical interface for storing and accessing translation data. Finally, the extraction of bilingual construction data other than those retrieved by means of automatic phrase generation is another one. Focus is set on verb valency with imperative and infinitive clauses as basic frame constructions.

References

Ahrenberg, L., M. Andersson and M. Merkel (1998), 'A simple hybrid aligner for generating lexical correspondences from parallel texts', in: *Proceedings of COLING-ACL'98*. Montreal, Canada. 29–35.

Ahrenberg, L., M. Merkel, K. Mühlenbock, D. Ridings, A. Sågvall Hein and J. Tiedemann (1998), 'Automatic processing of parallel corpora. A Swedish perspective'. Linköping: Electronic University Press. Also available at http://stp.ling.uu.se/~corpora/plug/.

Ahrenberg, L., M. Merkel, A. Sågvall Hein and J. Tiedemann (forthcoming) 'Evaluating LWA and UWA', PLUG deliverable 3A.1. Internal report.

Almqvist, I. and A. Sågvall Hein (1996), 'Defining ScaniaSwedish—a controlled language for truck maintenance', in: *Proceedings of the First International Workshop on Controlled Language Applications*. 26-27 March 1996. Centre for Computational Linguistics. Katholieke Universiteit Leuven.

Fung, P. and K. W. Church (1994), 'K-vec: A new approach for aligning parallel texts', in: *Proceedings from the 15th International Conference on Computational Linguistics* (Coling-94). Kyoto. 1096–1102.

Gale, W. A. and K.W. Church (1993), 'A program for aligning sentences in bilingual corpora', *Computational Linguistics*, 19(1): 75–102.

Melamed, D. I. (1997), 'A word-to-word model of translation equivalence', in: *Proceedings of the 35th Conference the Association for Computational Linguistics*, Madrid: Association for Machine Translation in the Americas.

Merkel, M. and L. Ahrenberg, (1999), 'Evaluating word alignment systems', PLUG Deliverable 2A.1. Internal report.

Merkel, M., M. Andersson and L.Ahrenberg (this volume), 'The PLUG Link Annotator—interactive construction of data from parallel corpora'. 77–94.

NUTEK (1999), 'Om maskinöversättning' [On Machine Translation].

Ridings, D. (1998), 'PEDANT. Parallel texts in Göteborg', *LEXIKOS* 8 (Afrilex-reeks/series 8:1998): 1–26.

Sågvall Hein, A. (1994), 'Preferences and linguistic choices in the Multra machine translation system', in: R. Eklund (ed.), *NODALIDA* '93 Proceedings of *'9:e Nordiska Datalingvistikdagarna'*, Stockholm 3-5 June 1993.

Sågvall Hein, A. (1997), 'Language control and machine translation', in: *Proceedings of the 7th International Conference on Theoretical and Methodological Issues in Machine Translation*. July 23-25, 1997. St. John's College, Santa Fe, New Mexico.

Smadja, F., K. R. McKeown and V. Hatzivassiloglou (1996), 'Translation collocations for bilingual lexicons: A statistical approach', *Computational Linguistics*, 22(1)*:* 1–38.

Tiedemann, J. (1998a), 'Extraction of translation equivalents from parallel corpora', in: *Proceedings of the 11th Nordic Conference on Computational Linguistics*, Copenhagen 28-29 January 1998 (NODALIDA'98), Center for Sprogteknologi, University of Copenhagen. 120–128.

Tiedemann, J. (1998b), 'Parallel corpora in Linköping, Uppsala and Göteborg (PLUG). Work package 1', Department of Linguistics, Uppsala University.

Tiedemann, J. (1999), 'Automatic construction of weighted string similarity measures', *Proceedings of EMNLP/VLC-99 (Joint Sigdat conference on empirical methods in natural language processing and very large corpora)*.

Tiedemann, J. (this volume), 'Uplug – a modular corpus tool for parallel corpora'. 107–122.

Wikholm, E., I. Maier, A. Östling and A. Sågvall Hein (1993), 'A multilingual dictionary of functional core phrases', Uppsala University, Department of Linguistics.

Appendix: Results of searching Scania 98 with the source search key ”tryck på”

id	link	source	target
sventscanTI1301	1-1	Välj fordons kategori "Lastbil 3-serien" och **tryck på** OK knappen.	Select the vehicle category "3 series truck" and press the OK button.
sventscanTI1304	1-1	Välj systemgrupp "ABS-system" och **tryck på** sök i fordon.	Select system group "ABS system" and press "Find in vehicle".
sventscanTI1529	1-1	Välj fordonskategori "Lastbil 3-serien" och **tryck på** OK-knappen.	Select the vehicle category "3 series truck" and press the OK button.
sventscanTI1532	1-1	Välj systemgrupp t ex "ABS-system" och **tryck på** Sök i fordonet.	Select the system group e.g. "ABS system" and press Find in vehicle.
sventscanSD9368	1-1	För lågt öppnings**tryck på** den insprutare som har nålrörelsegivare (vid avgasbromsning).	The control unit has sensed that resistance in the circuit between pins 32 and 17 has been too low or too high.
sventscanSD25851	1-1	Reglermodulen har känt ett kvarstående **tryck på** mer än 0,8 bar under pulstestet.	The control module has sensed a residual pressure of more than 0.8 bar during the pulse test.
sventscanSD25853	1-1	Felkoden bildas när det pulstest, som startar automatiskt när tändningen slås på, lämnar ett rest**tryck på** mer än 0,8 bar.	The fault code is generated when the pulse test, which starts automatically when the ignition is switched on, leaves a residual pressure of more than 0.8 bar.
sventscanSD25869	1-1	Reglermodulen har känt ett kvarstående **tryck på** mer än 0,8 bar under pulstestet.	The control module has sensed a residual pressure of more than 0.8 bar during the pulse test.
sventscanSD25872	1-1	Felkoden bildas när det pulstest, som startar automatiskt när tändningen slås på, lämnar ett rest**tryck på** mer än 0,8 bar.	The fault code is generated when the pulse test, which starts automatically when the ignition is switched on, leaves a residual pressure of more than 0.8 bar.

sventscanSD26184	1-1	Reglermodulen har känt ett kvarstående **tryck på** mer än 0,8 bar under pulstestet.	The control module has sensed a residual pressure of more than 0.8 bar during the pulse test.
sventscanSD26186	1-1	Felkoden bildas när det pulstest, som startar automatiskt när tändningen slås på, lämnar ett rest**tryck på** mer än 0,8 bar.	The fault code is generated when the pulse test, which starts automatically when the ignition is switched on, leaves a residual pressure of more than 0.8 bar.
sventscanSD26202	1-1	Reglermodulen har känt ett kvarstående **tryck på** mer än 0,8 bar under pulstestet.	The control module has sensed a residual pressure of more than 0.8 bar during the pulse test.
sventscanSD26205	1-1	Felkoden bildas när det pulstest, som startar automatiskt när tändningen slås på, lämnar ett rest**tryck på** mer än 0,8 bar.	The fault code is generated when the pulse test, which starts automatically when the ignition is switched on, leaves a residual pressure of more than 0.8 bar.
sventscanSD26517	1-1	Reglermodulen har känt ett kvarstående **tryck på** mer än 0,8 bar under pulstestet.	The control module has sensed a residual pressure of more than 0.8 bar during the pulse test.