On Portability of Resources for a Quick Ramp up of Multilingual MT of Patent Claims

Svetlana Sheremetyeva

LanA Consulting ApS
Jacobys Allé 23
DK-1806 Copenhagen Denmark
lanaconsult@mail.dk

Abstract

We describe a feasibility study on reusing the components of the unilingual authoring application AutoPat in a full-scale multilingual MT system APTrans, and explore to which extent MT knowledge can be ported from one language to another in the patent domain. We illustrate our findings on the example of English, Danish and French languages.

Introduction

Patents are a rich source of information about technological knowledge and a valuable tool in technology development. It is the area, which shows an increasing interest in high quality multilingual machine translation systems. To develop such systems requires rich knowledge resources (lexicons, grammar rules, world models), which nowadays must normally be painstakingly handcrafted from scratch for every language pair.

The idea to reduce development and maintenance costs, by sharing and reusing processing methods and knowledge has been in focus of researchers' attention for many years. For example, (Takeda, 1994) proposes portable knowledge sources for machine translation that consists of preference information on word sense, phrasal attachment, and word selection for translation. The basic idea of (Paul, 2001) is to devote efforts to the development of translation engines between the main linguistically different languages and to reuse the translation knowledge of these systems for translation into languages closely related to the target language. (Pinkham et al., 2001) describe the assembly of the French-English research MT system, which was constructed from a combination of pre-existing rule-based components and automatically created components.

A patent specific research in MT where the problem of portability is addressed by suggesting the constraint domain approach has been done for Russian to English by (Sheremetyeva and Nirenburg, 1999). Among the most recent attempts to reduce development cost by reusing pre-existing application components is a Japanese-English authoring patent system, which merges the English claim authoring system AutoPat (Sheremetyeva, 2003) and the Japanese machine translation application PC-Transfer (Neumann, 2005).

In this paper we present the results of further work on reusability of the AutoPat application. We describe a feasibility study on reusing the components of the existing unilingual authoring application in a full-scale multilingual MT system APTrans, and explore to which extent linguistic MT knowledge can be ported from one language to another in the patent domain. We illustrate our findings on the example of English, Danish and French in the frame of the APTrans architecture. Our discussion will mainly address the effort saving issues of augmenting the system with every new language-pair.

In what follows we shall first sketch the starting point of our research, the English patent claim authoring system AutoPat, we shall then describe the migration process from the unilingual AutoPat to the multilingual machine translation system APTrans followed by a worked out example for the three languages, - English, Danish and French. We shall also discuss other possibilities to use the APTrans architecture in machine translation.

AutoPat

AutoPat is a computer system for authoring patent claims in the English language. It consists of a technical knowledge elicitation module with an interactive user interface, lexicon, human input analysis module, content representation language, and generation module integrated with proofing tools (spelling, content and grammar checkers).

The knowledge base of the system includes a patent corpus-based English lexicon over a rich feature space, rules and knowledge representation language. AutoPat is a fully implemented product level application described in detail in (Sheremetyeva, 2003) and available at www.lanaconsult.com. We shall thus skip the AutoPat specification but rather concentrate on a re-engineering issue.

Development process: migration from unilingual authoring to multilingual MT.

Our goal is to find ways to speed up the development of a multilingual machine translation system, which can be specifically supported by domain constraints. Our multiyear R&D in the patent domain gave us a strong evidence of high lexical and structural similarity of patent claims in different languages. This inspired us to extrapolate "what is already there", - the knowledge base and program components of AutoPat, to another application, an MT system, and other languages.

Design

The first step in developing APTrans was to define a subset of the existing AutoPat components that will be the basis of the multilingual application and the extension it will need.

The modular architecture of AutoPat, which generates patent claims form content representations, suggested a transfer type MT architecture. All of the AutoPat components with the exception of the knowledge elicitation module can be reused for generation of the TL claim from the TL content representation. What is missing is an analyzing component, which could map raw claims into the AutoPat content representation format in a SL and a transfer module, which could convert a SL content representation into a TL content representation keeping the AutoPat format.

The knowledge base should be extended with multilingual MT lexicons, rules and heuristics. Other components that will definitely be needed are output post-editors for TLs.

To be a viable application that can be developed within a reasonable time a developer's environment for knowledge acquisition and maintenance should be an integral part of the application.

In our research the whole translation procedure was built "around" the existing AutoPat knowledge base and generator. We shall therefore first describe the reuse and customization of the lexicon, knowledge representation and generator and then show how the rest of the AP-Trans components were attached to them.

From the very start we programmed APTrans as a multilingual (not just bilingual) application, so that a new language can be easily integrated into the previously developed software.

Reuse and customization of existing components

Lexicon and feature space

The AutoPat lexicon (its vocabulary, entry format and feature space) is completely transferred to the APTrans application and used as a seed lexicon for lexical acquisition in other languages. We reused the approach to treat passive and active forms of verbs as different lexemes to simplify processing procedures.

Every entry following the English lexicon format is maximally defined as a tree of features:

SEM-CL [Language [POS [MORPH CASE_ROLE
FILLER PATTERN],

where

SEM_Cl - semantic class; POS - part of speech; MORPH – morphological features, such as number, gender, etc., and domain relevant wordforms; CASE_ROLEs, - a set of lexeme case roles such as *agent, theme, place, instrument*, etc; FILLERs – lexical categories that can fill case-role slots of a lexeme; PATTERNs - linking features, that code both the knowledge about co-occurrences of lexemes with their case-roles and the knowledge about their linear order in the claim text.

Every node in the APTrans tree of features inherits its values from its ancestor. The mechanism of inheritance works in such a way that, in general, most values are inherited from the closest ancestor unless it is blocked or overwritten.

What is not trivial and probably only possible in such a restricted domain as ours is that there is a significant cross-linguistic parallelism (portability) in the values of two features, - CASE-ROLEs, and PATTERNs.

In other words, the set of case-roles for crosslingually equivalent predicates (verbs) and the order of their realization in the claim text are essentially invariant across languages. It means that in our tree of features there is not only a traditional "vertical" inheritance from parents to children, but for certain sibling nodes there is also a "horizontal" cross linguistic value inheritance which saves a lot of effort in non-English lexical acquisition.

Content representation language

AutoTrans reuses the AutoPat claim content representation language on both SL and TL sides of the translation process.

The format of the claim content representation as a set of predicate templates is given in Figure 1, where "label" is a unique identifier of the elementary predicate-argument structure, "predicate-class" is a label of a semantic class, "predicate" is a string corresponding to a predicate from the system lexicon, "case-roles" are "ranked" according to the frequency of their cooccurrence with a certain predicate in the training corpus, "status" is a semantic status of a case-role, such as *agent, theme, place, instrument*, etc., and "value" is a string which fills a case-role.

Sentence::={ template){template}*
template::={label predicate-class predicate ((case-role)(case-role))*}
case-role::= (rank status value)
value::= phrase{(phrase(word tag)*)}*

Figure 1. A claim content representation format.

Generation module

The AutoPat generation module, which takes a TL set of templates as input is what APTrans profits most of. It is fully reused from AutoPat for the English TL and, as our experiments show so far, requires only a slight updating for a non-English TL.

The whole concept of AutoPat generation, its rules and algorithms were originally worked out for Russian, and they actually code the legal requirements to the claim structure, which are essentially the same all over the world. This gave us the idea to port the generation knowledge to the English AutoPat, where it is now used without any essential changes.

We repeated our exercise in APTrans and ported the generation rules, this time, from English to Danish and French. For both languages only a few rules were updated, mainly to cover TL subject-predicate agreement.

In those cases where updating the English generation rules for Danish or French required too much effort we left them unchanged, thus "programming" mistakes in the translation output. We found it easier to correct these predictable mistakes at a later stage of processing, by running a TL posteditor on the generator output.

Analyzer

It was natural to think of the APTrans analyzer as the component to output its parse in the format of the content representation language.

Trying to reuse the knowledge we have already acquired for the English AutoPat we started with the analyzer for the English language and built it "on top" of the AutoPat disambiguating tagger. A bottom-up heuristic parser with a recursive pattern matching technique was then added to recursively chunk longer phrases preserving their inner structure. It also marks the head of every noun phrase and "learns" its "singular/plural" feature.

The last analyzing procedure determines the dependency relations between the chunks and predicates, and puts these chunks as fillers into case-role slots in predicate/argument structures, thus defining their semantic status (Sheremetyeva, 2003).

The reuse of the AutoPat generator has the advantage of simplifying the analysis task by making it possible to skip the problems of determining a) the syntactic relations between the predicate and its arguments within every individual predicate structure (*microsyntax*), and b) the syntactic hierarchy of predicate/argument structures in the input claim text (*macrosyntax).*

The generator, as was mentioned above, has the microsyntactic and macrosyntactic knowledge about the template hierarchy and the order of the phrases within predicate templates coded in its rules and lexicon.

To test the compatibility of the analyzer and the generator we modeled a "translation" experiment within one (English) language, thus avoiding (for now) lexical transfer problems. Raw English claims were input into the analyzer, and parsed. The parse was input into the generator. The modules proved to be compatible and the results of such "translation" showed a reasonably small number of failures, mainly due to the incompleteness of analysis rules.

We then tried to port the English analysis knowledge to the analyzers for Danish and French, the experiments show so far that a great deal of English analysis rules in our domain and approach can also be reused, though, of course, language specificity requires customization (e.g., location of adjectives in French noun phrases, lexical clues, etc.).

Transfer module

The APTrans transfer module takes the analyzer output, - a SL set of predicate templates as input and outputs a set of TL predicate templates whose slots are filled with presumably perfectly translated TL phrases/case-role fillers.

The APTrans transfer is in fact a combination of interlingual and syntactic transfer. The interlingual transfer finds TL equivalents[1] for every predicate and keeps the predicate template slot structure unchanged (invariant). The syntactic transfer is responsible for the translation of case-role strings.

A "real" translation procedure is thus reduced to the phrase level which, though not without problems, is still much simpler than machine translation of a full patent claim, especially when, which is often the case, it runs for a page or so.

Translation of phrases is done in two runs. First all lexical items in the SL case-role fillers are simply looked up in the lexicon and substituted by the base forms of their TL equivalents.

The second run applies syntactic transfer rules to the case-role strings. These rules are responsible for syntactic restructuring and agreement in TL language phrases. Besides the knowledge in the TL lexicon the rule condition part relies on the knowledge about the case-role, the type of phrase to which the lexeme belongs and the tag history. The tag history is the knowledge about the tag (e.g., part-of-speech) of the equivalent lexeme in the SL, which might be different from that in the SL.

The rules for phrase translation are of course language dependent, but here again a certain amount of portability is possible. We first tried our approach on the English/Danish pair, - the first pair of phrase translation rules was written for the English to Danish direction. These rules mostly coverer some Danish morphology phenomena[2], and noun-article-adjective agreement in gender, definiteness and number.

In our experiments with the English to French translation we discovered that the left sides of agreement rules, which formulate the context for agreement, can in many cases be reused for the French language. The right sides of such rules, provided the reordering of adjectives is covered can to a certain extent be reused as well.

A worked example

Consider the following input claim text[3] in English to be translated in Danish and French:

A support for bearings comprising two connected half-shells provided with corresponding cavities adapted to form a seat for a bearing, characterized in that at least one of the cavities is shaped to form three radial raised portions for the contact of the bearing along corresponding imaginary lines parallel to the rotation axis of the bearing.

We illustrate this procedure on the example of translation a patent claim from English into French. The procedure for Danish is the same.

//A parsed output: English Predicate structures

```
Generic
(P1  Pgw  "comprising"
    1  Det1N2Prep3Np4 "A support  for bearings "
    2  Num5Pdc6Np7 "two connected half-shells ")

(P2p  Pdw  "provided"
    1  Num5Pdc6Np7 "two connected  half-shells  "
    2  Prepmn8Adjo9Np10 "with corresponding cavities")

(P3p  Pdg  "adapted"
    1  Adjo9Np10 "open corresponding  cavities "
    3  Infm11Pgvi12Det1N13 "to form a  seat ")
```

[1] A base form of TL predicate from the lexicon substitutes a SL predicate gloss. The TL predicate gloss can be changed in the generator according to the generation rules.

[2] For example, in the Danish language a definiteness of a noun is expressed morphologically: *a cup=en kop; the cup = koppen*

[3] For illustration we take a very short claim, normally the claim, still one sentence, can run for a page or so.

(P4 Pdgpr "for"
1 Det1N13 " a seat "
2 Det1N4 " a bearing ")

Difference

(P5p Pdvs "is shaped"
1 Qu14Qun15Detpl16Np10 "at least one of the cavities "
6 Infm11Pgvi12Num17Adjo18Pdo19Np20 "open to form three radial raised portions ")

(P6p Pdgpr "for"
1 Num17Adjo18Pdo19Np20 "three radial raised portions "
2 Detd16No21Prep22Detd16N4 "the contact of the bearing "
4 Prep23Adjo9Adjo24Np25 "along corresponding imaginary lines ")

(P7p Pdl "parallel"
1 Adjo9Adjo24Np25 "open corresponding imaginary lines
2 Prepmn11Detd16No26Prep22Detd16N4 "open to the rotation axis of the bearing ")

//French Predicate structures after BASE TRANSFER

Generic

(P1 W Pgw "comportant"
1Det1N35Prep17N6 "un soutien de roulement”
2Num41Pdc14Nfem19 "deux relié moitié-coquille ")

(P2p W Pdw "equipées"
1Num41Pdc14Nfem19 "deux relié moitié-coquille "
2 Prepmn42Adjo16Nfem7 "de correspondant cavité ")

(P3p G Pdg "adaptées"
1Adjo16Nfem7 "correspondant cavité close"
3 Prep39Pgv18Det2N34 " pour formant un siège")

(P4 G Pdg "pour"
1Det2N34 "un siège close"
2Det2N5 "un roulement")

Difference

(P5p V Pdv "formé"
1Qu4Qun23Detdm37Nfem7 "au moins un des le cavité close"
6 Prep39Pgv18Num38Adjo31Pdo32Nfem3 " pour formant trois radial augmentées partie ")

(P6p G Pdg "pour"
1Num38Adjo31Pdo32Nfem30 "trois radial augmentées partie
2Detdm36No15Prep22Detdm36N5 "le contact close de le roulement close"
4 Prep3Adjo16Adjo20Nfem21 " le long correspondant imaginaire ligne close")

(P7p L Pdl "parallèlles"
1Adjo16Adjo20Nfem21 "correspondant imaginaire ligne close"
2 Prepmn40Detdm36No33Prep22Detdm36N5 " à le axe de rotation de le roulement close”)

//French Predicate structures after RULE TRANSFER

Generic

(P1 Pgw "comportant"
1 Det1N2Prep3Np4 "un soutien de roulements"
2 Num5Nfemp7Pdcp6 "deux moitié-coquilles reliées")

(P2p Pd "equipées"
1 Num5Nfemp7Pdcp6 "deux moitié-coquilles reliées"
2 Prepmn8Nfemp10Adjfmp9 "de cavités correspondantes")

(P3p Pdg "adaptées"
1 Nfemp10Adjfmp9 "cavités correspondantes"
3 Prep11Pgvi12Det1N13 "pour former un siège")

(P4 Pdg "pour"
1 Det1N13 "un siège"
2 Det1N4 "un roulement")

Difference

(P5p Pdv "formé"
1 Qu14Qunfm15Nfemp10 "au moins une des cavités"
6 Prep11Pgvi12Num17Nfemp20Pdo19Adjfmp18 "pour former trois parties augmentées radiales")

(P6p Pdg "pour"
1 Num17Nfemp20Pdo19Adjfmp18 "trois parties augmentées radiales"
2 Detdm16No21 "le contact"
4 Prep22Detdm16N4 "de le roulement")

(P7p Pdl "parallèlles"
1 Detdpl0Nfemp24Adjfmp23Adjfmp9 "des lignes imaginaires correspondantes"
2 Prepmn11Detdm16No25 "à le axe de rotation")

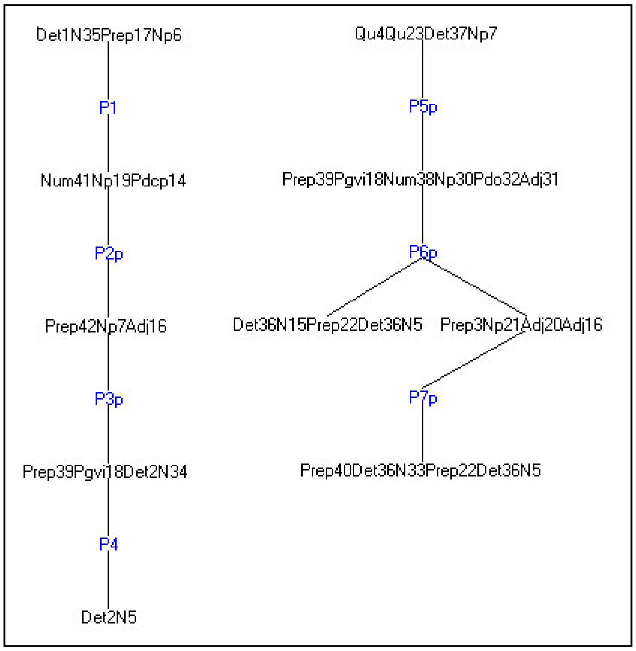
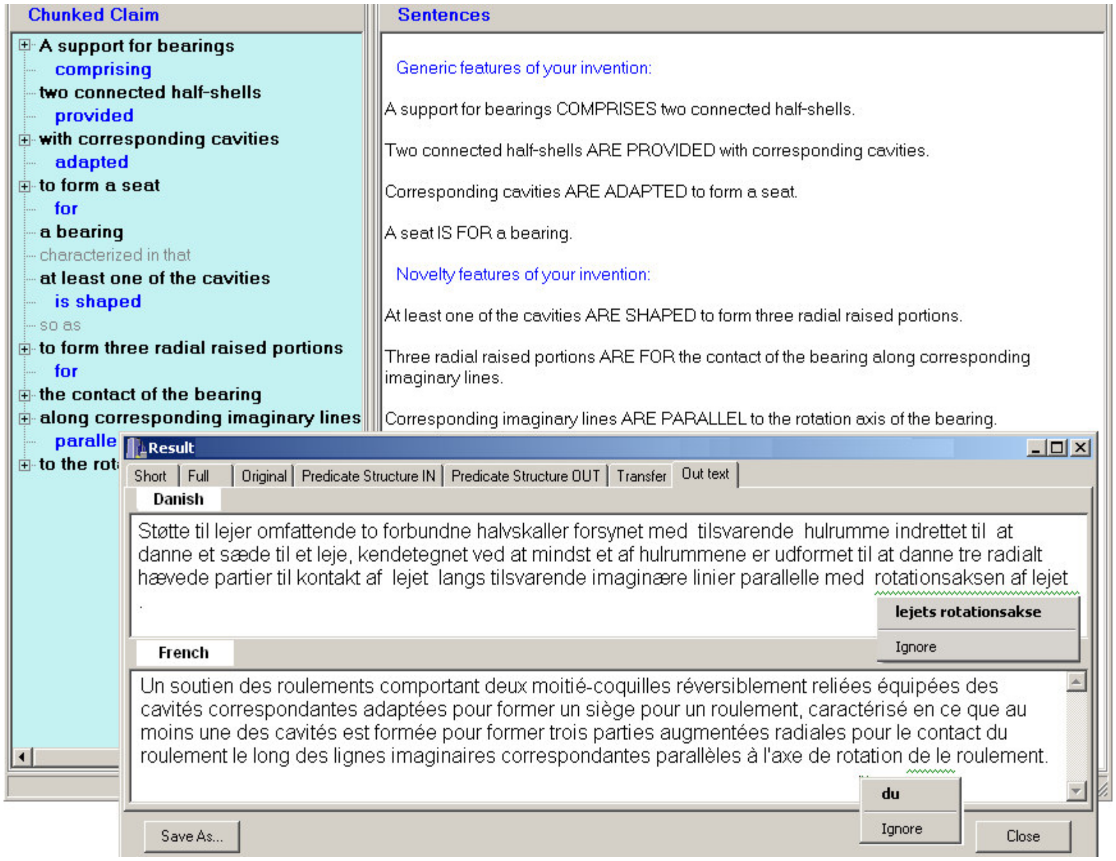


Figure 2. Trees built of the predicate templates by the generator.

French predicate structures after the RULE TRANSFER stage are input to the generator. All further operations are performed over strings of tags, which are substituted with the corresponding language phrases only after all the generation transformations are done. The input predicate templates are glued into trees following hard-coded language independent rules (See Figure 2). These trees following other set of generation rules, mainly universal, are linearised into a string of tags, which is further transformed to define the macrostructure and text cohesion of the TL French claims.

Figure 3 shows a screenshot of the professional user (e.g., translator) interface with the resulting APTrans translation from English into French and Danish. A trace of the postediting procedure is shown for every language.



A support for bearings
comprising
two connected half-shells
provided
with corresponding cavities
adapted
to form a seat
for
a bearing
characterized in that
at least one of the cavities
is shaped
so as
to form three radial raised portions
for
the contact of the bearing
along corresponding imaginary lines
paralle
to the rot

Generic features of your invention:

A support for bearings COMPRISES two connected half-shells.

Two connected half-shells ARE PROVIDED with corresponding cavities.

Corresponding cavities ARE ADAPTED to form a seat.

A seat IS FOR a bearing.

Novelty features of your invention:

At least one of the cavities ARE SHAPED to form three radial raised portions.

Three radial raised portions ARE FOR the contact of the bearing along corresponding imaginary lines.

Corresponding imaginary lines ARE PARALLEL to the rotation axis of the bearing.

Støtte til lejer omfattende to forbundne halvskaller forsynet med tilsvarende hulrumme indrettet til at danne et sæde til et leje, kendetegnet ved at mindst et af hulrummene er udformet til at danne tre radialt hævede partier til kontakt af lejet langs tilsvarende imaginære linier parallelle med rotationsaksen af lejet .

Un soutien des roulements comportant deux moitié-coquilles réversiblement reliées équipées des cavités correspondantes adaptées pour former un siège pour un roulement, caractérisé en ce que au moins une des cavités est formée pour former trois parties augmentées radiales pour le contact du roulement le long des lignes imaginaires correspondantes parallèles à l'axe de rotation de le roulement.

Figure 3. A screenshot of the APTrans user interface with an English claim translated in Danish and French.

The trace of the English claim analysis is shown to the user in the left pane of the background window.

The right pane of the background window shows simple sentences generated from the individual predicate templates. We kept this functionality from the original AutoPat generator for the user to check the correctness of the input claim analysis.

In case the simple sentences in the right pane are incorrect the user can interfere into the analysis procedure and through a special interface interactively correct the structure of the sentences thus correcting the analyzer output of predicate templates. This will result in a corrected translation.

Outsourcing MT

Reduction of the translation procedure to the machine translation on a phrase level opens another possibility for speeding up the multilingual translation development process: outsourcing phrase translation to a foreign MT system. We had a successful experience in trying this approach in a joint project on developing the Japanese-English patent authoring system[4], a patent claim generator in English from a Japanese-only interface. A Japanese user input the technical knowledge in his native language, which was further transformed by the system into a claim content representation in the AutoPat format with Japanese case-role fillers. The Japanese case-role fillers were separately translated from Japanese into English by the PC-Transfer MT system (see Neumann, 2005). The English strings were afterwards put back to the slots of predicate templates

[4] *The J-E patent system* ,Cross Language KK, Tokyo, Japan and LanA Consulting, Denmark, Copenhagen.

and input into the AutoPat Generator. As a result a full English translation of a Japanese claim was generated.

Performing MT by translating text segments smaller than sentence is getting into the focus of the MT research. (Bart et al., 2006) report on positive results achieved by reducing MT to a phrase level. In their experiment statistical techniques are used to decompose sentences into chunks, select the best translation of the chunks and recompose the translated chunks into a target language sentences.

Conclusions

In this paper we addressed the problem of saving on software development when building a family of NLP applications that share domain and task requirements. We illustrated the approach on the example of migrating from a system for authoring patent claims in English, AutoPat, to a multilingual machine translation system APTrans.

Though our research is a feasibility study we got a strong evidence that in the patent claim domain a noticeable economy of development effort could be achieved by porting linguistic machine translation knowledge from one language to another. We illustrated our findings on the example of English, Danish and French languages in the frame of the APTrans system architecture.

Due to the patent domain knowledge portability, as well as modularity of APTrans and the specificity of its components a foreign MT system can easily be integrated into the system architecture. This is a complementary way of speeding up the MT development.

We are planning to continue our research in both directions, - developing in-house machine translation resources and experimenting with foreign MT systems to integrate into APTrans those of them that show good results in their performance.

Bibliographical References

Bart. M., Mellebeek, K. Owczarzak, J.Van Genabith & A.Way. (2006). Multi-Engine Machine Translation by Recursive Sentence Decomposition. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, Cambridge, August 2006.

Neumann Ch. (2005). A Human-Aided Machine Translation System for Japanese-English Patent Translation. Proceedings of the Workshop on Patent Translation in Conjunction with MT Summit, Phuket, Thailand, September 16

Paul M. (2001): Translation Knowledge Recycling for Related Languages. *Proceedings of MT Summit VIII18-22* September. Santiago de Compostela, Galicia, Spain.

Pinkham J., M.Corston-Oliver, M. Smets & M.Pettenaro. (2001). Rapid assembly of a large-scale French-English MT system. *Proceedings of MT Summit VIII18-22*. September. Santiago de Compostela, Galicia, Spain.

Takeda K. (1994). Portable Knowledge Sources for Machine Translation. Proceedings of COLING 1994, 15th International Conference on Computational Linguistics, August 5-9. Kyoto, Japan.

Sheremetyteva, S. and S. Nirenburg. (1999). Interactive MT As Support For Non-Native Language Authoring. *Proceedings of the MT Summit VII. September 13-17, 1999, Singapore*.

Sheremetyteva S. (2003a). Towards Designing Natural Language Interfaces. Proceedings of the 4th International Conference “Computational Linguistics and Intelligent Text Processing” Mexico City, Mexico, February 16-22.

Sheremetyeva S. (2003b). Natural Language Analysis of Patent Claims. *Proceedings of the workshop “Patent Corpus Processing” in conjunction with 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003*), Sapporo. Japan, July 7-12.