PETRA – the Personal Embedded Translation and Reading Assistant

Werner Winiwarter
Faculty of Computer Science, University of Vienna
Liebiggasse 4
A-1010 Vienna, Austria
werner.winiwarter@univie.ac.at

Abstract

In this paper we present PETRA, a Personal Embedded Translation and Reading Assistant, which assists German-speaking language students in reading and translating Japanese documents. PETRA is fully embedded into Microsoft Word so that the students can invoke all its features directly from within the text editor. The translation rules are learnt incrementally from translation examples provided by the students, therefore, they can customize the translations according to their personal preferences. This type of language learning environment encourages a bidirectional knowledge transfer, which fosters a lively interaction and makes the learning process more interesting and entertaining.

1 Introduction

In general, reading and studying online material available via the Web represents an excellent way to improve the fluency in a foreign language because new terminology and new grammatical constructs can be learnt within their natural context with comparatively little effort. However, for Japanese documents this approach to language acquisition soon turns into a frustrating experience for the language student due to the complexity of the Japanese writing system.

Japanese writing is a mixture of Chinese characters called *kanji* and the two syllabaries *hiragana* and *katakana*. Whereas the syllabaries are relatively easy to learn with only 46 different characters each, there are several thousand, mostly quite complex kanji characters. Today 1,945 kanji are used in the basic education curriculum and in publications for the general public.

Another severe problem with trying to read a Japanese text is that Japanese writing does not use word delimiters (such as space characters) so that the student has to guess the word boundaries. If the student reads a sentence he cannot understand, he must first find out where an unknown word starts before he can consult a dictionary. However, the student can only use a bilingual dictionary if he knows the correct pronunciation of the word. Unfortunately, kanji can have several pronunciations or *readings* depending on the context. Therefore, quite often the student must consult a kanji dictionary in which kanji are categorized according to 214 basic elements or *radicals.* All this makes reading and translating Japanese documents a cumbersome and tedious process.

Thus, the aim of our research is to support German-speaking language students at the University of Vienna with a *Personal Embedded Translation and Reading Assistant (PETRA)*. We have implemented a language learning environment using Amzi! Prolog, which provides an application programming interface to Visual Basic to enable the full integration into Microsoft Word. Therefore, the language student can use the text editor to work with any Japanese document and access all the features of the reading and translation assistant from within Microsoft Word.

The rest of the paper is structured as follows. In Sect. 2 we present the features of the language learning environment. Section 3 gives a brief overview of the machine translation component, for a more detailed technical description we refer to (Winiwarter, 2004). Finally, we close the paper with some concluding remarks and an outlook on future work.

2 System Description

PETRA can be activated directly from any position within a text document by using a simple shortcut. The PETRA workspace is opened in a separate window and the sentence surrounding the active cursor position is extracted and displayed in the new window. Figure 1 shows an example of a Japanese document on the history of books. The student started PETRA for the fifth Japanese sentence, asked for the pronunciation (“youhishi”) and the meaning (“parchment”) of the first word, and finally for the translation of the whole sentence into German (“Parchment was made of sheep or goat skins, vellum of calf skins”).

The implemented functionality of our *reading assistant* includes the correct segmentation of Japanese sentences and the dictionary lookup of pronunciations and German meanings. The dic-

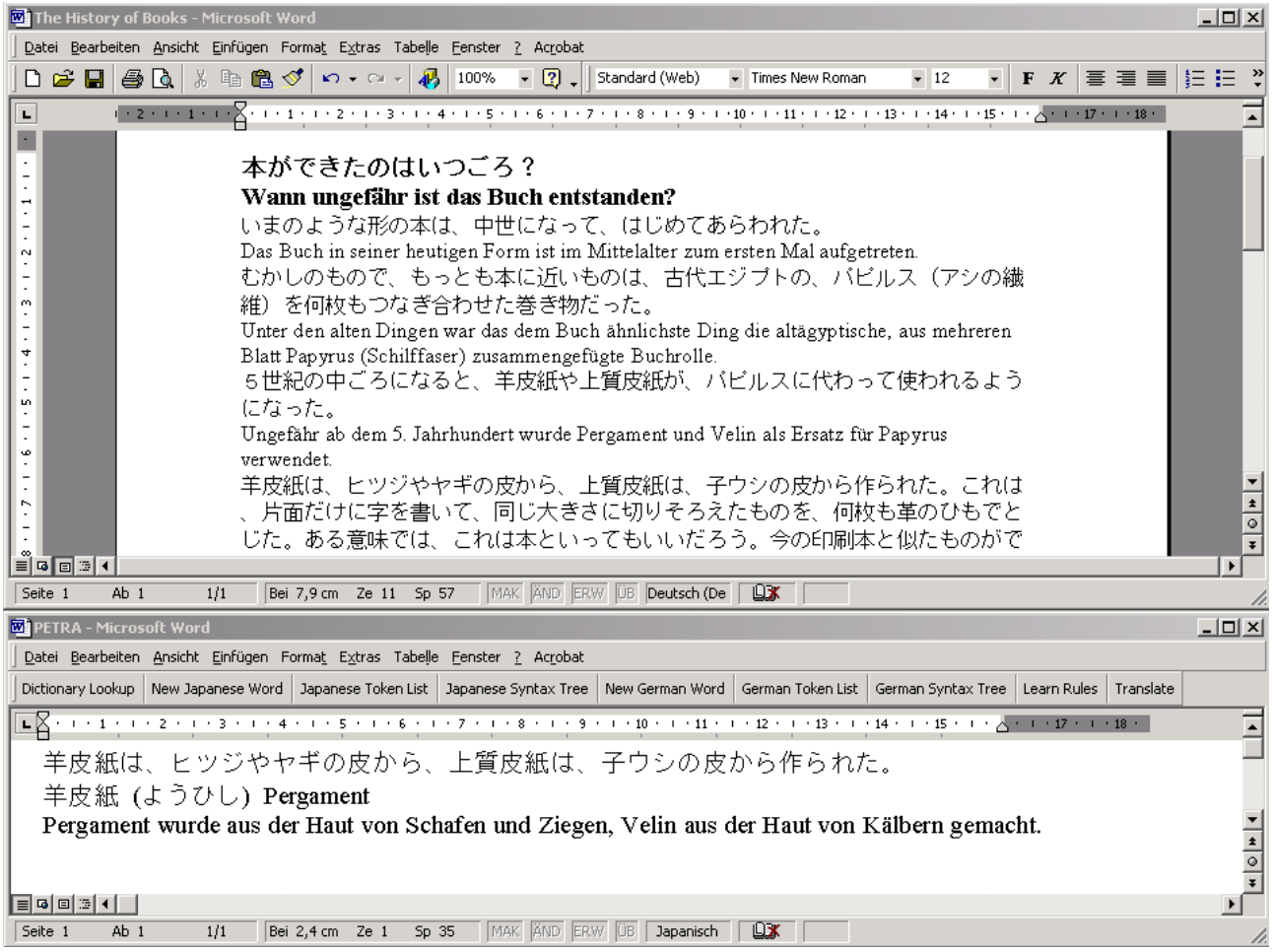

Figure 1: Example of user interface

tionary lookup is based on the Japanese–German dictionary WaDokuJT (www.wadoku.de) containing 190,251 entries. Since there exists a large number of conjugations for Japanese verbs and adjectives, an important feature is *stemming* to find the dictionary form for conjugated word forms. Finally, we also provide a user-friendly interface to add new entries to the dictionary.

The *translation assistant* automatically translates Japanese sentences into German. If the language student is not satisfied with the translation result, he can simply correct the translation and activate the adaptive learning module, which results in an update of the translation rule base. This way the students can fully personalize the language learning environment according to their individual needs and preferences.

This incremental improvement of the translation quality of PETRA encourages a bidirectional knowledge transfer between the language student and the system. PETRA supplies valuable information to the student and the student makes use of this information to solve the translation task at hand, which involves the student in an active role during the whole interaction with PETRA.

In addition to the final translation result the student can also request the display of token lists and syntax trees for Japanese and German sentences, which are generated as intermediate results by the machine translation component. The *token lists* are the result of the tokenization modules and show the correct segmentation of a sentence into word tokens associated with their part-of-speech tags. Finally, the *syntax trees* are produced by the parsing modules and provide a structural representation of a sentence by indicating the syntactic constituents. Figure 2 shows the token list and syntax tree for the example sentence from Fig. 1.

3 Machine Translation

Research on machine translation has a long tradition, for good overviews see (Hutchins and Somers, 1992), (Newton, 1992), or (Hutchins, 2003a). The state of the art in machine translation is that there exist quite good solutions for narrow application domains with a limited vocabulary and concept space. It is the general opinion that fully automatic high quality machine translation without any limitations on the subject and without human intervention is far beyond the scope of today's machine translation technology and there is serious doubt that it will be ever possible in the future (Hutchins, 2003b). It is very disappointing to have

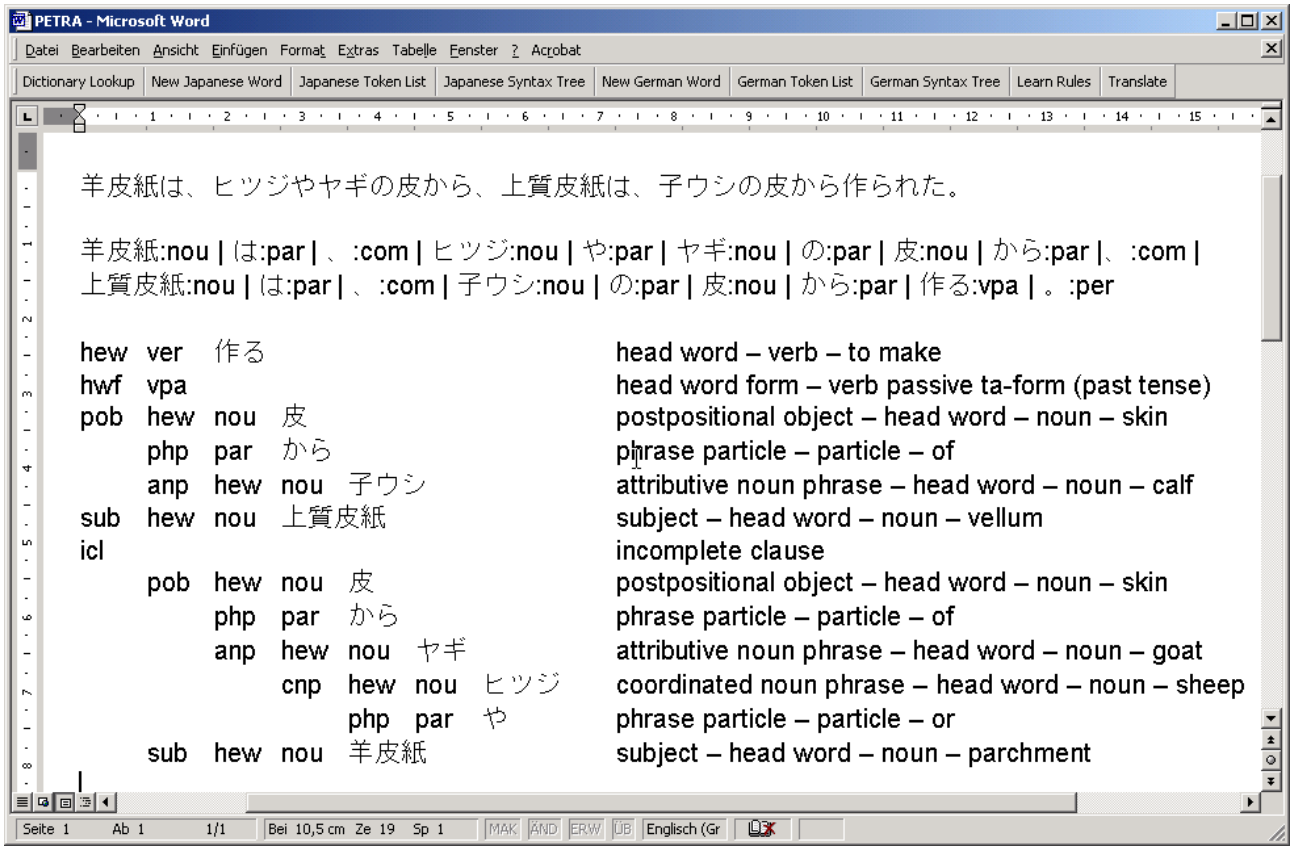

Figure 2: Example of token list and syntax tree

to notice that the translation quality has not much improved in the last 10 years (Somers, 2003). Figure 3 shows the results of an entertaining experiment with some of the Japanese machine translation programs that are freely available on the Web, where we tried to obtain translations for the example sentence from Fig. 1. Only the first system offers translations into both German and English, all others are restricted to English.

Japanese sentence:
羊皮紙は、ヒツジやヤギの皮から、上質皮紙は、子ウシの皮から作られた。

Machine translation by WorldLingo
(www.worldlingo.com/products_services/worldlingo_translator.html):

Was das Pergament anbetrifft, von der Haut der Schafe und??, was das obere Qualitätshautpapier anbetrifft, es wurde gebildet vom Kasten.

As for the parchment, from the skin of the sheeps and ヤギ, as for the upper quality skin paper, it was formed from the box.

Machine translation by Excite (www.excite.co.jp/world/url/):
Leather paper of fine quality was made from the skin of HITSUJI or a goat for 羊皮紙 from the skin of a child cow.

Machine translation by TransLand
(www.brother.co.jp/jp/honyaku/demo/index.html):
As for the sheep skin paper, good skin paper was made from the skin of the child cow from the skin of the sheep and the goat.

Machine translation by iTranslator
(itranslator.mendez.com/BGSX/BGSXeng_us-EntryPage.htm):
For parchment, with a skin of a sheep and a goat, fine quality skin paper was made with a skin of a child bull.

Figure 3: Example of machine translation systems

One main obstacle on the way to achieving better translation quality is caused by the fact that most of the existing translation systems are not able to learn from their mistakes. They contain large static rule bases with limited coverage, which have been assembled with huge intellectual effort. All the valuable feedback provided by users through post-editing translation results is usually lost for future translations.

Whereas there exist several bilingual corpora for Japanese–English, which can be exploited for example-based machine translation, e.g. see (Brockett et al, 2002), (Watanabe et al, 2002), or (Yamada, 2002), we had no such resources available for the language pair Japanese–German.

Thus, for PETRA we have developed a transfer-based architecture in which the transfer rules are learnt directly from translation examples provided by the language student. This means that the system does not rely on an inflexible collection of handcrafted rules, but, on the contrary, the language student can fully customize the system.

Figure 4 shows the architecture of PETRA's machine translation component. Each sentence pair is first analyzed by the *tokenizer* modules, which produce the correct segmentations into word tokens associated with their part-of-speech tags. Because of the missing delimiters we have to treat a Japanese sentence as a single string and retrieve all left substrings – including concatenations of stems and endings for conjugated words – from the lexicon to identify the next correct word.

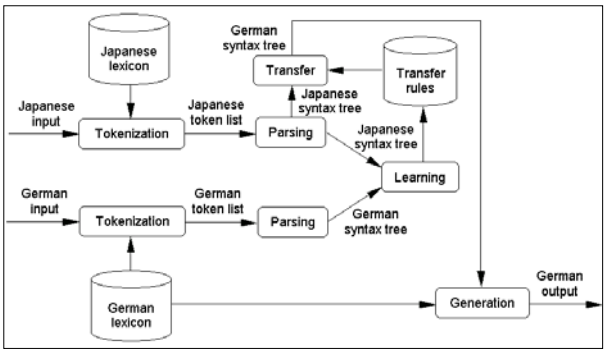

Figure 4: Machine translation architecture

Both Japanese and German token list are then transformed into syntax trees by the *parsing modules*. For specifying the syntax trees, we use a flexible and robust representation by modeling a sentence as a set of *constituents*. Each constituent is either a *simple constituent* (feature or word) or a *complex constituent*, which consists again of a set of subconstituents. We have implemented several generic predicates to manipulate complex constituents, e.g. to find, insert, remove, or replace subconstituents. This compact representation enables an efficient application of transfer rules during translation because the transfer rules can change the trees incrementally without having to consider any ordering information.

For displaying the syntax trees to the language students we have developed one generic *display module*, which can handle both Japanese and German syntax trees as well as mixed representations caused by missing coverage of the transfer rule base. This way we can show the limitations of the translation component to the language student who can fix them with an update of the rule base.

The *learning module* traverses both syntax trees and derives new transfer rules, which are added to the rule base. For that purpose we use generic predicates for the simultaneous navigation in two complex constituents. We start looking for mappings at the top level of the sentence before searching for corresponding constituents and continuing the search for finer-grained rules recursively.

For translating a Japanese sentence, we activate the *transfer module*. It incrementally applies the transfer rules stored in the rule base to the Japanese syntax tree. One rule only changes certain parts of a constituent into the German counterpart, other parts are left unchanged to be transformed later on. Thus, our transfer algorithm deals efficiently with a mixture of Japanese and German, which gradually turns into a correct German syntax tree.

Finally, the *generation module* traverses the German syntax tree in the correct order and produces a list of word tokens along the way. This list is then transformed into a single character string by inserting spaces where necessary. The main difficulty is the generation of the correct inflected word forms at the surface level. The required syntactic information is partly encoded in the syntax tree (e.g. number or tense) and partly derived from the German lexicon (e.g. gender of nouns).

4 Conclusion

We have finished the implementation of the system and are now in the process of building the transfer rule base with the help of several language students from the University of Vienna. So far, the feedback from the students has been very positive. For some, PETRA has already become an invaluable companion throughout their language studies.

In addition to constantly extending the coverage of our rule base, future work will also concentrate on a detailed evaluation study to measure the impact of PETRA on language instruction along the three dimensions engagement, effectiveness, and viability.

References

C. Brockett et al. 2002. English-Japanese example-based machine translation using abstract semantic representations. *Proc. of the COLING-2002 Workshop on Machine Translation in Asia.*

J. Hutchins. 2003a. Has machine translation improved? Some historical comparisons. *Proc. of the 9th MT Summit*, pages 181-188.

J. Hutchins. 2003b. Machine translation and computer-based translation tools: What's available and how it's used. In J. M. Bravo (ed). *A New Spectrum of Translation Studies*, University of Valladolid.

J. Hutchins and H. Somers. 1992. *An Introduction to Machine Translation.* Academic Press.

J. Newton (ed). 1992. *Computers in Translation: A Practical Appraisal.* Routledge.

H. Somers (ed). 2003. *Computers and Translation: A Translator's Guide.* John Benjamins.

T. Watanabe et al. 2002. Statistical machine translation based on hierarchical phrase alignment. *Proc. of the 9th Intl. Conference on Theoretical and Methodological Issues in Machine Translation*, pages 188-198.

W. Winiwarter. 2004. Incremental learning of transfer rules for customized machine translation. *Proc. of the 15th Intl. Conference on Applications of Declarative Programming and Knowledge Management*, pages 183-192.

K. Yamada. 2002. *A Syntax-Based Statistical Translation Model*. PhD Thesis, University of Southern California.